On Fast Leverage Score Sampling and Optimal Learning

Alessandro Rudi*[,1], Daniele Calandriello*[,2], Luigi Carratino[3] and Lorenzo Rosasco[2,3]

* Equal contribution [1] INRIA – Sierra-project team & École Normale Supérieure, Paris, France
[2] LCSL – Istituto Italiano di Tecnologia, Genoa, Italy & Massachusetts Institute of Technology, Cambridge, USA
[3] DIBRIS – Università degli Studi di Genova, Genoa, Italy

Neural Information Processing Systems Foundation

Motivation & Contribution

A successful approach used to scale kernel methods to large datasets is to
(1) find a **small** set of Nyström centres that **accurately approximates** the kernel matrix,
(2) run a **fast kernel regression solver** exploiting the approximation.
The second step is currently much faster than the first, and the **bottleneck** of this pipeline is **constructing an accurate set** of Nyström centres.
We propose **BLESS**, to remove this bottleneck. It is currently the fastest Nyström center selection algorithm with provable accuracy guarantees.

Leverage Score Sampling

Given a **dataset** $D = \{x_1, \dots, x_n\}$, a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,
let the kernel matrix $\widehat{K} \in \mathbb{R}^{n \times n}$, be such that $\widehat{K}_{ij} = k(x_i, x_j)$
define the **leverage score** [1] as

LS $\quad \ell(i, \lambda) = \left(\widehat{K} \left(\widehat{K} + \lambda n I \right)^{-1} \right)_{ii} \quad$ with $\quad i \in [n]$

Goal: sampling $J = \{i_1, \dots, i_M\} \subset \{1, 2, \dots, n\} = [n]$ according to $\ell(i, \lambda)$

	Uniform	LS
$\lvert J \rvert$	$d_\infty(\lambda)$	$d_{\text{eff}}(\lambda)$

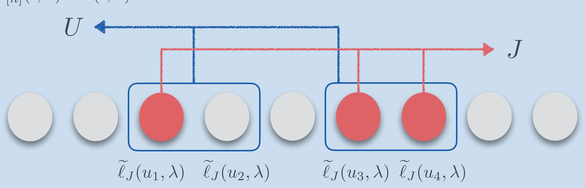
with $\quad d_\infty(\lambda) = n \max_{i=1,\dots,n} \ell(i, \lambda)$, $\quad d_{\text{eff}}(\lambda) = \sum_{i=1}^{n} \ell(i, \lambda)$,

Note: $d_{\text{eff}}(\lambda) \le \frac{1}{\lambda}$

Computing **LS** is **too expensive**. We need an approximation [3]:

Approx-LS $\quad \widetilde{\ell}_J(i, \lambda) = \frac{1}{\lambda n} \left(\widehat{K}_{ii} - \widehat{K}_{Ji}^\top \left(\widehat{K}_{JJ} + \lambda n A \right)^{-1} \widehat{K}_{Ji} \right) \quad$ with $\quad i \in [n]$

Note that $\widetilde{\ell}_{[n]}(i, \lambda) = \ell(i, \lambda)$



$L_J(U, \lambda) = \{\widetilde{\ell}_J(u_1, \lambda), \dots, \widetilde{\ell}_J(u_R, \lambda)\} \qquad L_J(U, \lambda) \mapsto J'$

Chicken-and-Egg problem: need J to compute **Approx-LS**, need **Approx-LS** to compute J

Previous Algorithms

Two-Pass sampling [1]: Given $J_1 \sim \text{Uniform}$ and $U_2 = [n]$ $\quad L_{J_1}(U_2, \lambda) \mapsto J_2$

Recursive-RLS [2]: Given $U_1 \subset U_2 \subset U_3 = [n]$ with $U_1, U_2 \sim \text{Uniform}$

Initialize $J_1 = U_1$ $\quad L_{J_1}(U_2, \lambda) \mapsto J_2 \quad L_{J_2}(U_3, \lambda) \mapsto J_3$

SQUEAK [3]: Given U_1, U_2, U_3 partition of $[n]$ $\quad$ Initialize $J_1 = U_1$

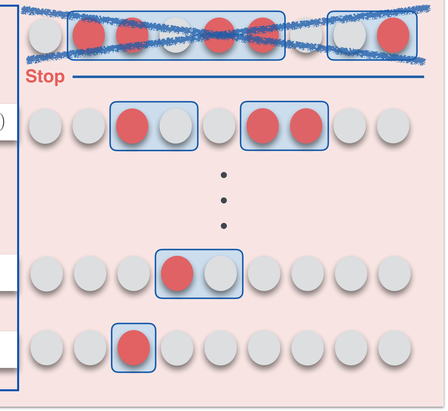
$L_{J_1 \cup U_2}(J_1 \cup U_2, \lambda) \mapsto J_2 \qquad L_{J_2 \cup U_3}(J_2 \cup U_3, \lambda) \mapsto J_3$

BLESS

$\lambda_k = \frac{1}{\sqrt{n}} \quad U_k \sim \text{Unif}\left(\frac{1}{\lambda_k}\right) \quad J_k = L_{J_{k-1}}(U_k, \lambda_k)$

$\lambda_2 \approx \frac{\lambda_1}{2} \quad U_1 \sim \text{Unif}\left(\frac{1}{\lambda_1}\right) \quad J_2 = L_{J_1}(U_2, \lambda_2)$

$\lambda_1 \approx 1 \quad U_2 \sim \text{Unif}\left(\frac{1}{\lambda_2}\right) \quad J_1 = U_1$



Theoretical Analisys

Theorem: Let $n \in \mathbb{N}$, $\lambda > 0$ and $\delta \in (0, 1]$. Given $t > 0$ and $H \in \mathbb{N}$, $(\lambda_h)_{h=1}^H \sim (1/2^{h-1})_{h=1}^H$, when $(J_h, A_h)_{h=1}^H$ are computed using **BLESS**, with probability at least $1 - \delta$:

(a). $\quad \frac{1}{1+t} \ell(i, \lambda_h) \le \widetilde{\ell}_{J_h}(i, \lambda_h) \le (1+t) \ell(i, \lambda_h), \quad \forall i \in [n], h \in [H]$

(b). $\quad |J_h| \le \mathcal{O}(d_{\text{eff}}(\lambda_h) \log(Hn/\delta)/t^2)), \quad \forall h \in [H]$

Computational complexity:

Cost of step h of **BLESS**: $\quad |J_h|^3 + |U_h||J_h|^2 \le d_{\text{eff}}(\lambda_h)^3 + \frac{1}{\lambda_h} d_{\text{eff}}(\lambda_h)^2 \le 2 \frac{1}{\lambda_h} d_{\text{eff}}(\lambda_h)^2$

Note: Set $\quad H = \log_2\left(\frac{1}{\lambda}\right) \longrightarrow \lambda = \lambda_H = \frac{1}{2^H} \longrightarrow \sum_{h=1}^H \frac{1}{\lambda_h} = \frac{1}{\lambda} \sum_{h=1}^H \frac{1}{2^{h-1}} \le \frac{2}{\lambda}$

Overall runtime: $4 d_{\text{eff}}(\lambda)^2 / \lambda$

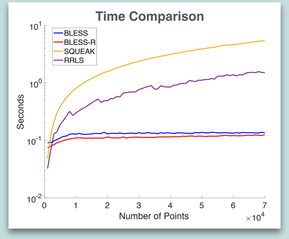
Improves over existing algorithms:

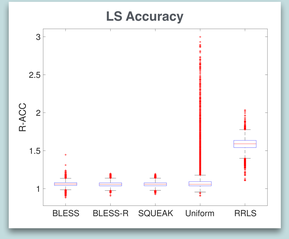
- compute whole regularization path $(\lambda_h)_{h=1}^H$ **for free**
  $\hookrightarrow$ J_h accurate at scale λ_h
- **early stop** at desired regularization $\lambda = \lambda_H$
  $\hookrightarrow$ never compute all n $\ell(i, \lambda)$ (breaks linear bottleneck)

Algorithm	Runtime	$\lvert J \rvert$
Uniform sampling	–	$1/\lambda$
Exact RLS sampl.	n^3	$d_{\text{eff}}(\lambda)$
Two-Pass sampling	n/λ^2	$d_{\text{eff}}(\lambda)$
Recursive-RLS	$n d_{\text{eff}}(\lambda)^2$	$d_{\text{eff}}(\lambda)$
SQUEAK	$n d_{\text{eff}}(\lambda)^2$	$d_{\text{eff}}(\lambda)$
BLESS	$\mathbf{1/\lambda\ d_{\text{eff}}(\lambda)^2}$	$d_{\text{eff}}(\lambda)$

Key new tools:

- control error $L_{J_h}([n], \lambda_h) \longrightarrow L_{J_h}(U_h, \lambda_h)$ $\qquad$ **hint**: $|U_h| \sim d_\infty(\lambda_h) \le 1/\lambda_h$ is sufficient
- control error $L_{J_h}(U_h, \lambda_h) \longrightarrow L_{J_h}(U_{h+1}, \lambda_{h+1})$ $\qquad$ **hint**: $|J_h| \sim \frac{\lambda_h}{\lambda_{h+1}} d_{\text{eff}}(\lambda_h)$ is sufficient





Learning Setting

Given $(x_i, y_i)_{i=1}^n$ i.i.d. samples from ρ, minimize the **expected risk**

$$\min_{f \in \mathcal{H}} \mathcal{E}(f) \qquad \mathcal{E}(f) = \int (y - f(x))^2 \, d\rho(x, y)$$

RKHS of kernel k $\qquad$ unknown

FALKON-BLESS

Model: $\quad \widehat{f}_{\lambda, J}(x) = \sum_{j \in J} K(\widetilde{x}_j, x) \alpha_j$

subset of D chosen according to **LS**

computed with **BLESS**

Learning: $\quad \alpha = (K_{nJ}^\top K_{nJ} + \lambda K_{JJ})^{-1} K_{nJ}^\top y$

+ Conjugate gradient **+**

$$BB^\top = \left(\frac{n}{M} K_{JJ} A^{-1} K_{JJ} + \lambda K_{JJ} \right)^{-1}$$

Statistical Results

Theorem: Let $n \in \mathbb{N}$, $\lambda > 0$ and $\delta \in (0, 1]$. Given $t \ge \log(n)$ when $n \ge d_{\text{eff}}(\lambda)$, with probability at least $1 - \delta$:

$$\mathcal{E}(\widehat{f}_{\lambda, J}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \lesssim \frac{d_{\text{eff}}(\lambda)}{n} \log(1/\delta) + \lambda + e^{-t}$$

In particular, when $d_{\text{eff}}(\lambda) \lesssim \lambda^{-1/\alpha}$, for $\alpha \ge 1$, by selecting $\lambda_* = n^{-\alpha/(\alpha+1)}$

$$\mathcal{R}(\widehat{f}_{\lambda_*, J}) \lesssim n^{-\frac{\alpha}{\alpha+1}}$$

FALKON-UNIFORM	$d_\infty(\lambda_*) + n d_\infty(\lambda_*)$	$\lesssim$	$n^{\frac{\alpha}{\alpha+1}} + n \cdot n^{\frac{\alpha}{\alpha+1}}$
FALKON-SQUEAK/RRLS	$n d_{\text{eff}}(\lambda_*)^2 + n d_{\text{eff}}(\lambda_*)$	$\lesssim$	$n \cdot n^{\frac{2}{\alpha+1}} + n \cdot n^{\frac{1}{\alpha+1}}$
FALKON-BLESS	$d_{\text{eff}}(\lambda_*)^2/\lambda_* + n d_{\text{eff}}(\lambda_*)$	$\lesssim$	$n \cdot n^{\frac{1}{\alpha+1}} + n \cdot n^{\frac{1}{\alpha+1}}$

Experiments

Datasets: • SUSY: $n \approx 5 \cdot 10^6$ • HIGGS: $n \approx 10^7$

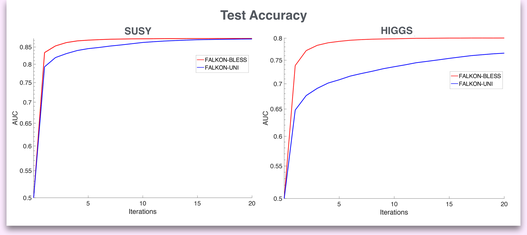
Much **faster convergence**

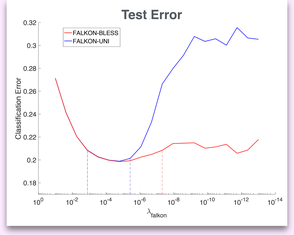
(1s + 10m vs 12s + 2m)

uniform $\quad$ **BLESS** $\quad$ **FALKON [4]**

(1s + 5h vs 1.5m + 1.4h)

Greater **stability** with respect to hyper-parameters





Limited accuracy improvement when using LS sampling over uniform. **Data too easy?**

References

[1] A. Alaoui and M. Mahoney. **Fast randomized kernel methods with statistical guarantees.** NeurIPS, 2015.
[2] C. Musco and C. Musco. **Recursive Sampling for the Nyström Method.** NeurIPS, 2017.
[3] D. Calandriello, A. Lazaric and M. Valko. **Distributed sequential sampling for kernel matrix approximation.** AISTATS, 2017.
[4] A. Rudi, L. Carratino and L. Rosasco. **FALKON: An Optimal Large Scale Kernel Method.** NeurIPS, 2017.