Analysis of Nyström Method with Sequential Ridge Leverage Score Sampling

Daniele Calandriello, Alessandro Lazaric, Michal Valko

Inria informatics mathematics

Motivation

- Kernel regression is *versatile* and *accurate*
- Strong accuracy guarantees but *poor scalability*

$\mathcal{O}(n^3)$ *time* $\mathcal{O}(n^2)$ *space* (n number of samples)

- *Current limitation:* Many approximate schemes are either **not scalable** or **not accurate**

⇒ **We propose an incremental approximation scheme for kernel regression with** ***complexity and error guarantees*** **depending on the** ***kernel structure***

Kernel Ridge Regression (KRR)

The setting (fixed-design)

- Dataset $\mathcal{D} = \{\mathbf{x}_t, y_t\}_{t=1}^n$
  - **arbitrary** $\mathbf{x}_t \in \mathcal{X}$
  - $y_t = f^*(\mathbf{x}_t) + \eta_t$
- Kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- Kernel matrix $\mathbf{K}_t \in \mathbb{R}^{t \times t}$, with $[\mathbf{K}_t]_{i,j} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j), i, j \le t$

Kernel regression

- Objective (after t samples)

$$\widehat{\mathbf{w}}_t = \arg\min_{\mathbf{w}} \|\mathbf{y}_t - \mathbf{K}_t\mathbf{w}\|^2 + \boldsymbol{\mu}\|\mathbf{w}\|^2.$$

- Closed-form solution

$$\widehat{\mathbf{w}}_t = (\mathbf{K}_t + \mu\mathbf{I})^{-1}\mathbf{y}_t$$

- On-sample risk

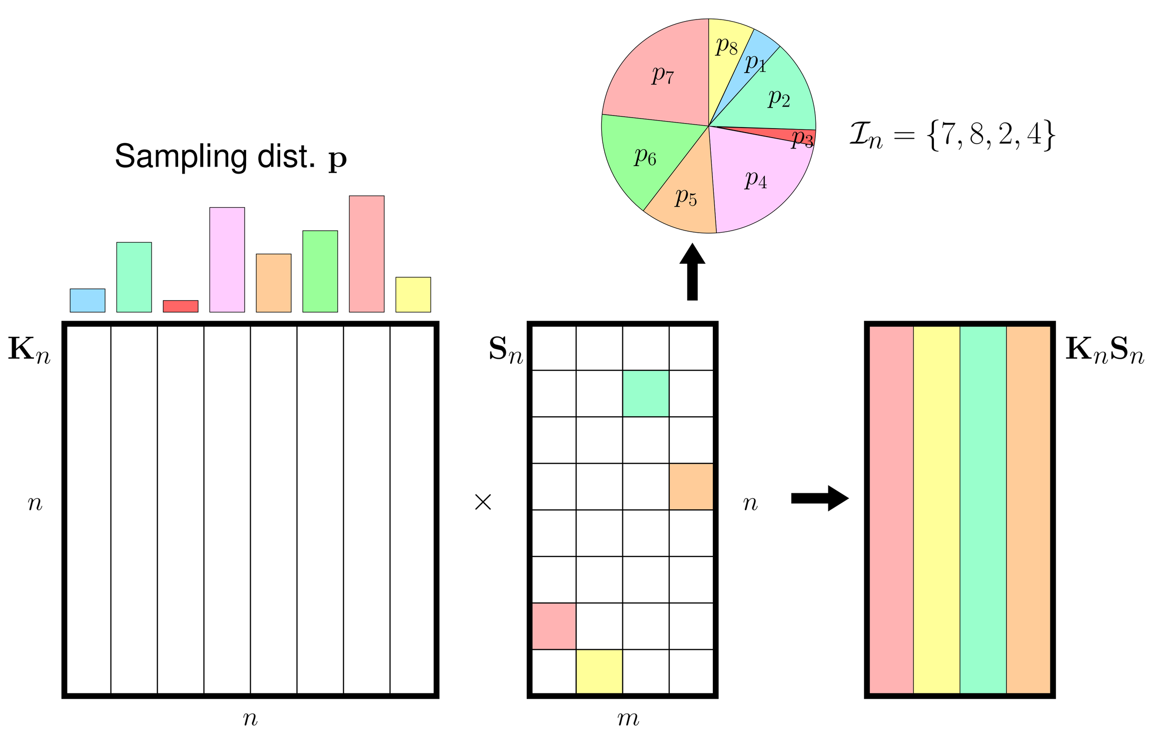
$$\mathcal{R}(\widehat{\mathbf{w}}_t) = \mathbb{E}_\eta\left[\|\mathbf{f}_t^* - \mathbf{K}_t\widehat{\mathbf{w}}_t\|^2\right]$$

Nyström Approximation

Subsampling

1 Select a subset (dictionary) $\mathcal{I}_n$ of $\boldsymbol{m}$ representative samples

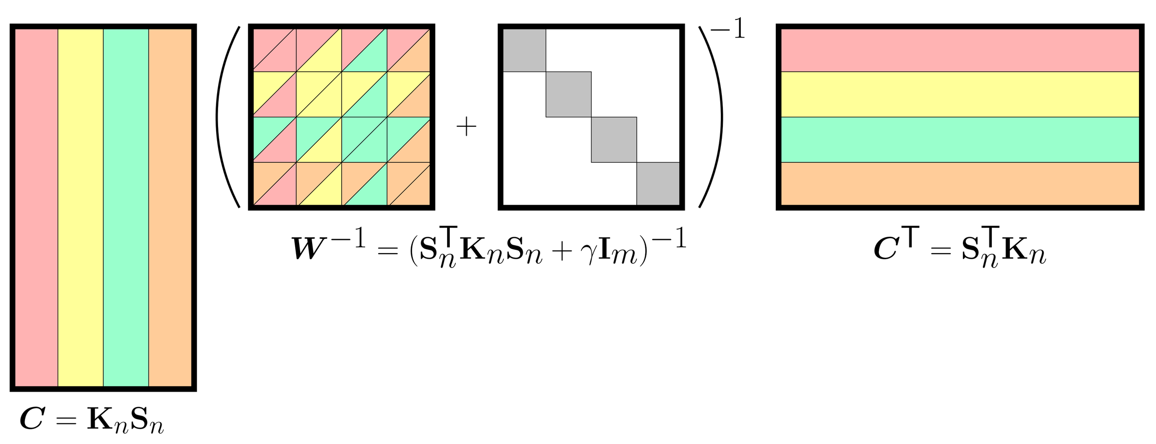
2 Constructs a sparse matrix $\mathbf{S}_n$ to select and reweight the columns associated with the points in $\mathcal{I}_n$



Low-Rank Approximation

3 Compute approximate, low-rank matrix $\widetilde{\mathbf{K}}_n = \mathbf{C}\mathbf{W}^{-1}\mathbf{C}^\mathsf{T}$ as

$$\widetilde{\mathbf{K}}_n = \mathbf{C}\mathbf{W}^{-1}\mathbf{C}^\mathsf{T} = \mathbf{K}_n\mathbf{S}_n(\mathbf{S}_n^\mathsf{T}\mathbf{K}_n\mathbf{S}_n + \boldsymbol{\gamma}\mathbf{I}_m)^{-1}\mathbf{S}_n^\mathsf{T}\mathbf{K}_n$$



Efficient Solution

4 Compute approximate solution

$$\widetilde{\mathbf{w}}_n = (\widetilde{\mathbf{K}}_n + \mu\mathbf{I})^{-1}\mathbf{y}_n = \frac{1}{\mu}\left(\mathbf{y}_n - \mathbf{C}\left(\mathbf{C}^\mathsf{T}\mathbf{C} + \mu\mathbf{W}\right)^{-1}\mathbf{C}^\mathsf{T}\mathbf{y}_n\right)$$

Scalability now depends on $\boldsymbol{m}$

Space: $\cancel{\mathcal{O}(n^2)} \Rightarrow \mathcal{O}(nm)$, Time: $\cancel{\mathcal{O}(n^3)} \Rightarrow \mathcal{O}(nm^2 + m^3)$

Problems:

? How to choose the sampling distribution?

? How to chose m?

References

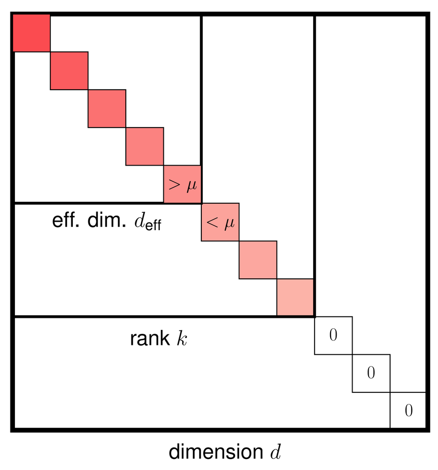
[Alaoui and Mahoney(2015)] Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel methods with statistical guarantees. In *Neural Information Processing Systems*, 2015.

[Bach(2013)] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *International Conference on Learning Theory*, 2013.

[Pachocki(2016)] Jakub Pachocki. Analysis of resparsification. *arXiv preprint arXiv:1605.08194*, 2016.

[Rudi et al.(2015)] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Neural Information Processing Systems*, 2015.

Kernel Ridge Leverage Scores (RLS) Sampling for KRR



Definition 1. *Given a kernel matrix* $\mathbf{K}_n \in \mathbb{R}^{n \times n}$*, define*

γ-ridge leverage score $\tau_{i,n}(\gamma) = \mathbf{k}_{i,n}^\mathsf{T}(\mathbf{K}_n + \gamma\mathbf{I}_m)^{-1}\mathbf{e}_{i,n}$ (1)

effective dimension $d_{\text{eff}}(\gamma)_n = \sum_{i=1}^n \tau_{i,n}(\gamma) = \operatorname{Tr}\left(\mathbf{K}_n(\mathbf{K}_n + \gamma\mathbf{I}_n)^{-1}\right)$ (2)

sampling distribution $[\mathbf{p}_n]_i = p_{i,n} = \frac{\tau_{i,n}(\gamma)}{\sum_{j=1}^n \tau_{i,n}(\gamma)} = \frac{\tau_{i,n}}{d_{\text{eff}}(\gamma)_n}$ (3)

Proposition 1 (Alaoui, Mahoney, 2015)**.** *Let $\boldsymbol{\varepsilon}$ be the accuracy, δ the confidence. If the regularized Nystrom approximation $\widetilde{\mathbf{K}}_n$ is computed using the sampling distribution $\{p_{i,t}\}$, and at least*

$$m \ge \left(\frac{2\boldsymbol{d_{\text{eff}}(\gamma)_n}}{\boldsymbol{\varepsilon^2}}\right)\log\left(\frac{n}{\delta}\right)$$

columns, then with probability $1 - \delta$

$$\mathbf{0} \preceq \mathbf{K_n} - \widetilde{\mathbf{K}}_{\boldsymbol{n}} \preceq \frac{\gamma}{1-\varepsilon}\mathbf{I}_n, \quad \mathcal{R}(\widetilde{\mathbf{w}}_n) \le \left(1 + \frac{\gamma}{\mu}\frac{1}{1-\varepsilon}\right)^2\mathcal{R}(\widehat{\mathbf{w}}_n)$$

Intuitively: $\tau_{i,n}$ *sensitivity* of prediction on point $\mathbf{x}_i$
$\Rightarrow \widehat{y}_{i,n} = \mathbf{e}_i^\mathsf{T}(\mathbf{K}_n\widehat{\mathbf{w}}_n) = \mathbf{e}_i^\mathsf{T}\mathbf{K}_n(\mathbf{K}_n + \mu\mathbf{I})^{-1}\mathbf{y}_n$

Pros: $+$ m scales with the effective dimension
$+$ the risk for $\widetilde{\mathbf{w}}_n$ is *almost* the same as for the exact solution

Cons: - computing $\tau_{i,n}(\mu)$ is as difficult as solving the original problem
- the probabilities need be **recomputed at any new sample** (=multipass)

Incremental Estimates of RLS and Effective Dimension

For any column i in $\mathcal{I}_t$ and $\mathbf{k}_{t+1}$ compute the ridge leverage score estimator ($\alpha = \frac{2-\varepsilon}{1-\varepsilon}$)

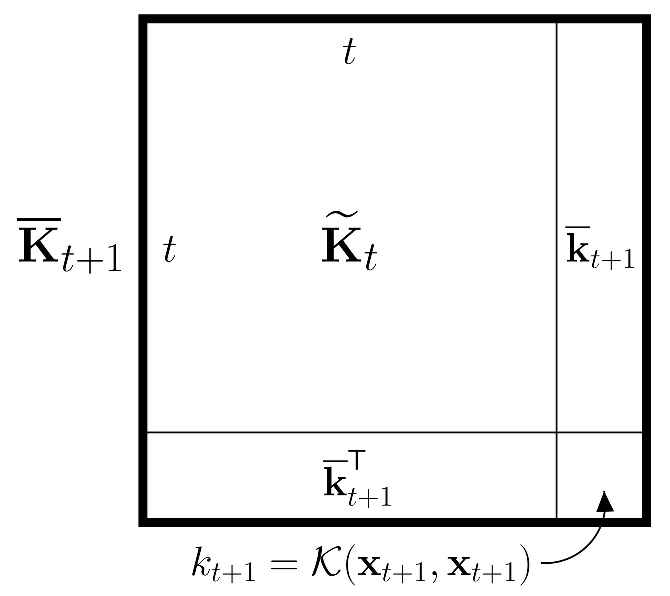
$$\widetilde{\tau}_{i,t+1} = \frac{1}{\alpha\gamma}\left(k_{i,i} - \mathbf{k}_{i,t+1}^\mathsf{T}\left(\widetilde{\mathbf{K}}_{t+1} + \alpha\gamma\mathbf{I}\right)^{-1}\mathbf{k}_{i,t+1}\right)$$

- $\widetilde{\tau}_{i,t+1} = \mathbf{e}_i^\mathsf{T}\widetilde{\mathbf{K}}_t(\widetilde{\mathbf{K}}_t + \gamma\mathbf{I})^{-1}\mathbf{e}_i$ would fail
- $\widetilde{\tau}_{i,t+1}$ is computed only for columns stored in $\mathcal{I}_t$ (accurate)
- $\widetilde{\tau}_{i,t+1}$ can be computed in a space/time efficient way
- α trades off accuracy of the estimator and space/time cost

Compute the effective dimension estimator $\widetilde{d}_{\text{eff}}(\gamma)_{t+1} = \widetilde{d}_{\text{eff}}(\gamma)_t + \alpha\widetilde{\Delta}_t$ with

$$\widetilde{\Delta}_t = \frac{\left(k_{t+1} - \overline{\mathbf{k}}_{t+1}^\mathsf{T}(\widetilde{\mathbf{K}}_t + \alpha\gamma\mathbf{I})^{-1}\overline{\mathbf{k}}_{t+1} - \frac{(1-\varepsilon)^2}{4}\gamma\overline{\mathbf{k}}_{t+1}^\mathsf{T}(\widetilde{\mathbf{K}}_t + \gamma\mathbf{I})^{-2}\overline{\mathbf{k}}_{t+1}\right)}{k_{t+1} + \gamma - \overline{\mathbf{k}}_{t+1}^\mathsf{T}(\widetilde{\mathbf{K}}_t + \alpha\gamma\mathbf{I})^{-1}\overline{\mathbf{k}}_{t+1}}$$

- $\widetilde{d}_{\text{eff}}(\gamma)_{t+1} = \sum_{i=1}^{t+1}\widetilde{\tau}_{i,t+1}$ requires $\widetilde{\tau}_{i,t+1}$ for $i \notin \mathcal{I}_t$ (not accurate)
- $\widetilde{\Delta}_t$ captures the interaction between the new and past samples
- $\widetilde{\Delta}_t$ requires approximating "second order" terms for which first order reconstruction guarantees ($\mathbf{0} \preceq \mathbf{K}_t - \widetilde{\mathbf{K}}_t \preceq \frac{\gamma}{1-\varepsilon}\mathbf{I}$) are not enough



Lemma 1. *Let ε be the accuracy and $\rho = \lambda_{\max}(\mathbf{K}_n)/\gamma$ a soft condition number. If after t samples $\widetilde{\mathbf{K}}_t$ is such that $\mathbf{0} \preceq \mathbf{K}_t - \widetilde{\mathbf{K}}_t \preceq \frac{\gamma}{1-\varepsilon}\mathbf{I}$, then for $\alpha = \frac{2-\varepsilon}{1-\varepsilon}$ and $\beta = \left(\frac{2-\varepsilon}{1-\varepsilon}\right)^2(1+\rho)$, the estimators satisfy for any $i \in \{\mathcal{I}_t \cup t+1\}$*

$$\frac{1}{\boldsymbol{\alpha}}\tau_{i,t+1}(\gamma) \le \widetilde{\tau}_{i,t+1} \le \mathbf{1}\cdot\tau_{i,t+1}(\gamma), \qquad \mathbf{1}\cdot d_{\text{eff}}(\gamma)_{t+1} \le \widetilde{d}_{\text{eff}}(\gamma)_{t+1} \le \boldsymbol{\beta}d_{\text{eff}}(\gamma)_{t+1}.$$

and the estimated probabilities satisfy

$$\frac{\mathbf{1}}{\boldsymbol{\alpha\beta}}p_{i,t+1} \le \widetilde{p}_{i,t+1} \le \mathbf{1}\cdot p_{i,t+1}$$

INK-Estimate

INK-Estimate
Input: Dataset $\mathcal{D}$, regularization γ, sampling budget $\overline{q}$
Output: $\widetilde{\mathbf{K}}_n$, $\mathbf{S}_n$
Initialize $\mathcal{I}_0$ as empty, $\widetilde{p}_{1,0} = 1, b_{1,0} = 1$, budget $\overline{q}$
for $t = 0, \dots, n-1$ **do**
  Receive new column $\overline{\mathbf{k}}_{t+1}$ and scalar k_{t+1}
  Compute *approximate leverage scores* $\{\widetilde{\tau}_{i,t+1} : i \in \mathcal{I}_t \cup \{t+1\}\}$
  Compute *approximate effective dimension* $\widetilde{d}_{\text{eff}}(\gamma)_{t+1}$
  Set $\widetilde{p}_{i,t+1} = \min\{\widetilde{\tau}_{i,t+1}/\widetilde{d}_{\text{eff}}(\gamma)_{t+1}, \widetilde{p}_{i,t}\}$
  $\mathcal{I}_{t+1}, \mathbf{b}_{t+1}$ = Shrink-Expand($\mathcal{I}_t, \widetilde{\mathbf{p}}_{t+1}, \mathbf{b}_t, \overline{q}$)
  Compute $\mathbf{S}_{t+1}$ using $\mathcal{I}_{t+1}$ and weights $\sqrt{b_{i,t+1}}$
  Compute $\widetilde{\mathbf{K}}_{t+1}$ using $\mathbf{S}_{t+1}$
end for
Return $\widetilde{\mathbf{K}}_n$ and $\mathbf{S}_n$

Theorem 1. *Let ε be the desired accuracy and $\rho = \lambda_{\max}(\mathbf{K}_n)/\gamma$ a soft condition number. If* INK-Estimate *is run with*

$$\overline{q} \ge \left(\frac{28\alpha\beta d_{\text{eff}}(\gamma)_t}{\varepsilon^2}\right)\log\left(\frac{4t}{\delta}\right),$$

then the approximate kernel solution $\widetilde{w}_n$ satisfies

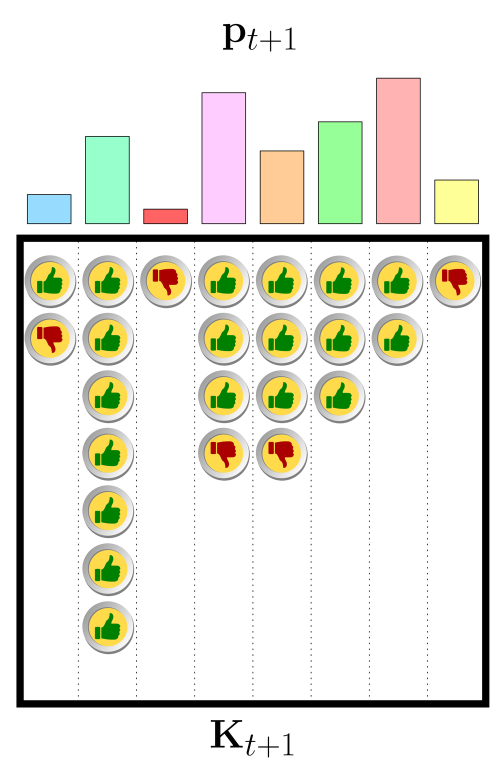
$$\mathcal{R}(\widetilde{w}_n) \le \left(1 + \frac{\gamma}{\mu}\frac{1}{1-\varepsilon}\right)^2\mathcal{R}(\widehat{w}_n)$$

and INK-Estimate *runs in at most*

$\mathcal{O}(n\overline{q}) \le \widetilde{\mathcal{O}}(\boldsymbol{n\rho d_{\text{eff}}(\gamma)_n})$ *space,*

$\mathcal{O}(n^2\overline{q}^2 + n\overline{q}^3) \le \widetilde{\mathcal{O}}(\boldsymbol{n^2\rho^2 d_{\text{eff}}(\gamma)_n^2})$ *time*

Shrink-Expand (Pachocki, 2016)
Input: $\mathcal{I}_t$, $\{(\widetilde{p}_{i,t+1}, b_{i,t}) : i \in \mathcal{I}_t\}$, $\widetilde{p}_{t+1,t+1}$, $\overline{q}$
Output: $\mathcal{I}_{t+1}$, the set of all columns with $b_{i,t+1} \neq 0$
$b_{i,t+1} = b_{i,t}$ for all $i \in [t]$, $b_{t+1,t+1} = 1$
for all $i \in \{1, \dots, t\} : b_{i,t} \neq 0$ **do** ▷Shrink
  while $b_{i,t+1}\widetilde{p}_{i,t+1} \le 1/\overline{q}$ **do**
    Sample a random Bernoulli $\mathcal{B}\left(\frac{b_{i,t+1}}{b_{i,t+1}+1}\right)$
    On success set $b_{i,t+1} = b_{i,t+1} + 1$
    On failure set $b_{i,t+1} = 0$, **break**
  end while
end for
while $b_{t+1,t+1}\widetilde{p}_{t+1,t+1} \le 1/\overline{q}$ **do** ▷Expand
  Sample a random Bernoulli $\mathcal{B}\left(\frac{b_{t+1,t+1}}{b_{t+1,t+1}+1}\right)$
  On success set $b_{t+1,t+1} = b_{t+1,t+1} + 1$
  On failure set $b_{t+1,t+1} = 0$, **break**
end while



Lemma 2. *For any kernel matrix $\mathbf{K}_t$ at time t, and its bordering $\mathbf{K}_{t+1}$ at time $t+1$,*

$$\frac{\tau_{i,t+1}}{d_{\text{eff}\,t+1}} = \boldsymbol{p_{i,t+1}} \le \boldsymbol{p_{i,t}} = \frac{\tau_{i,t+1}}{d_{\text{eff}\,t}}.$$

	Time	Space	Acc. loss	Inc.
Exact	n^3	n^2	1	/
Bach'13	$\frac{nd_{\max}{}^2 + d_{\max}{}^3}{\varepsilon}$	$\frac{nd_{\max}}{\varepsilon}$	$(1+4\varepsilon)$	No
A&M'15	$n(\text{space})^2$	$\left(\frac{\lambda_{\min} + n\mu\varepsilon}{\lambda_{\min} - n\mu\varepsilon}\right)nd_{\text{eff}} + \frac{\operatorname{Tr}(\mathbf{K}_n)}{\mu\varepsilon}$	$(1+2\varepsilon)^2$	No
Ink-Est	$\frac{\rho^2 n^2 d_{\text{eff}}{}^2}{\varepsilon}$	$\frac{\rho n d_{\text{eff}}}{\varepsilon}$	$(1+2\varepsilon)^2$	*Yes*

Pros:
+ Accuracy ***and*** space/time guarantees
+ In the worst case, only $\sqrt{n}$ space overhead (wrt exact method)
+ Anytime risk guarantees

Cons:
- The time complexity is not fully satisfactory
- The current formulation of the estimators is not "fully" incremental

Open questions:
? Removing the dependency on ρ
? Random design (Rudi et al., 2015)
? Online learning