Pack only the essentials: Adaptive dictionary learning for kernel ridge regression

Daniele Calandriello, Alessandro Lazaric, Michal Valko

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 $[\mathbf{K}_t]_{i,j} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j), i, j \leq t$

Kernel regression

- Objective (after t samples)

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

- Closed-form solution

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

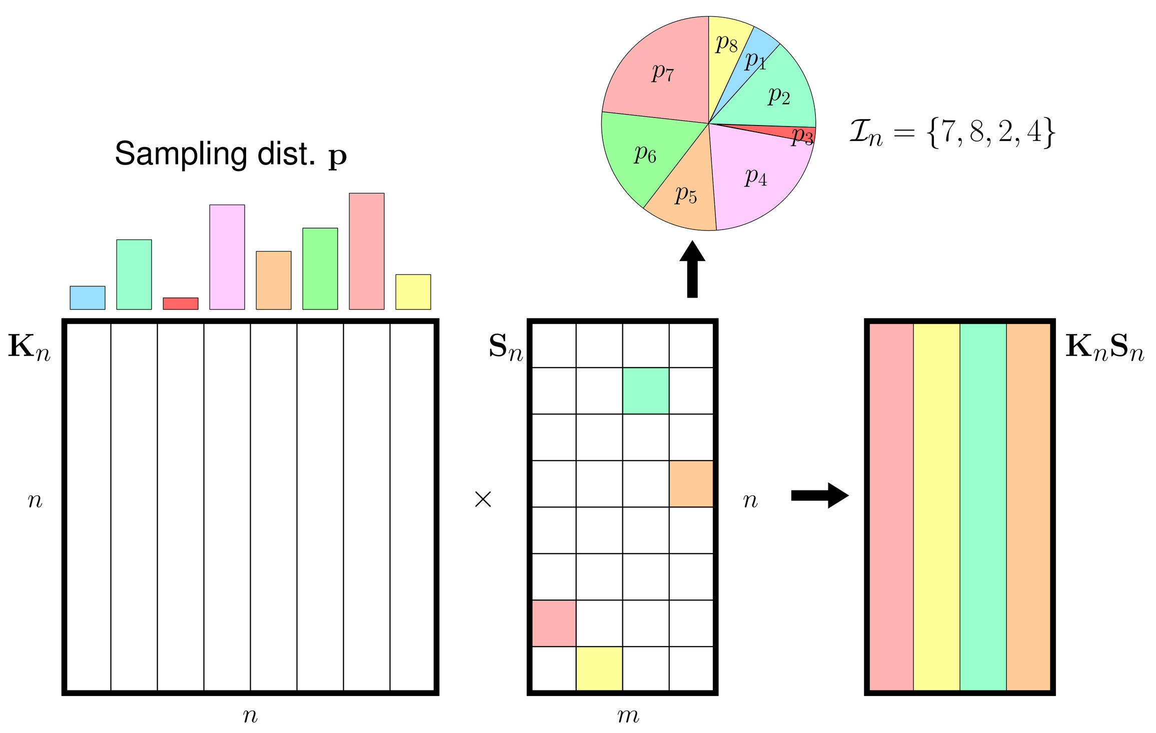
- On-sample risk

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

Nyström Approximation

Subsampling

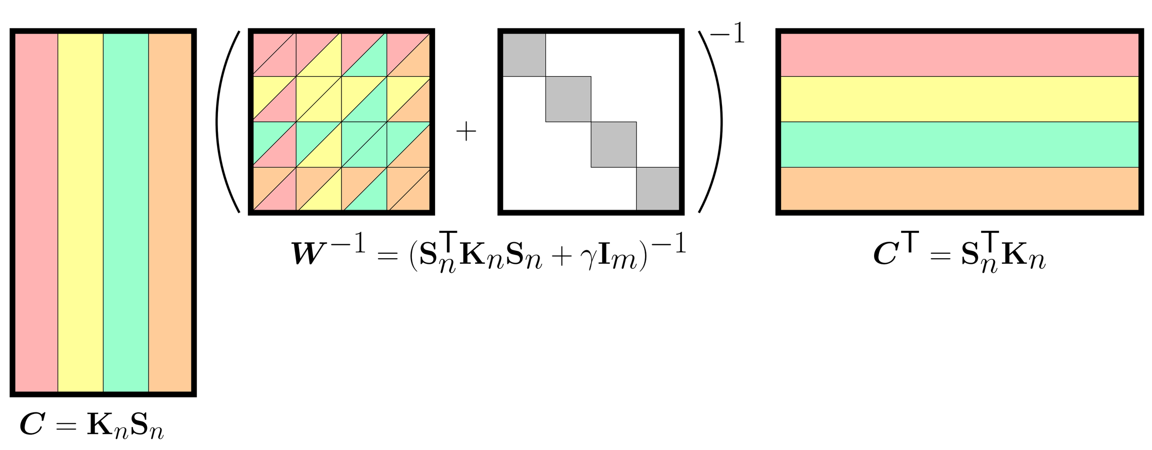
1 Select a subset (dictionary) $\mathcal{I}_n$ of m representative samples
2 Constructs a sparse matrix $\mathbf{S}_n$ to select and reweight the columns associated with the points in $\mathcal{I}_n$



Low-Rank Approximation

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

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



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 m

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

Problems:

? How to choose the sampling distribution?
? How to choose m?

References

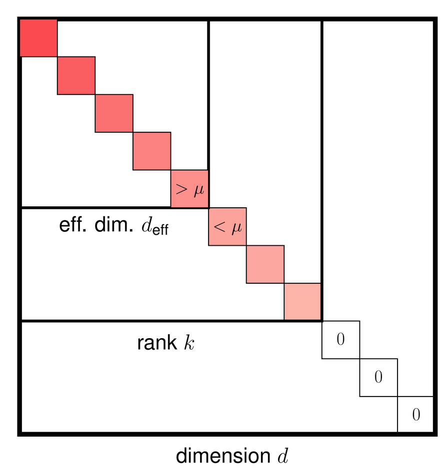
[Alaoui and Mahoney(2015)] A. El Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. In *NIPS*, 2015.

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

[Calandriello et al.(2016)] D. Calandriello, A. Lazaric, and M. Valko. Analysis of NystrÃűm method with sequential ridge leverage scores. In *UAI*, 2016.

[Rudi et al.(2015)] A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *NIPS*, 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

$$\tau_{n,i}(\gamma) = \mathbf{e}_{n,i}\mathbf{K}_n^\mathsf{T}(\mathbf{K}_n + \gamma\mathbf{I}_n)^{-1}\mathbf{e}_{n,i} = \phi(\mathbf{x}_i)^\mathsf{T}(\phi(\mathbf{X}_n)\phi(\mathbf{X}_n)^\mathsf{T} + \gamma\mathbf{I})^{-1}\phi(\mathbf{x}_i) \quad (1)$$

effective dimension

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

sampling distribution

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

Proposition 1 (Alaoui, Mahoney, 2015). *Let ε 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*

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

columns, then with probability $1 - \delta$

$$0 \preceq \mathbf{K}_n - \widetilde{\mathbf{K}}_n \preceq \frac{\gamma}{1-\varepsilon}\mathbf{I}_n, \qquad \mathcal{R}(\widetilde{\mathbf{w}}_n) \leq \left(1 + \frac{\gamma}{\mu}\frac{1}{1-\varepsilon}\right)^2\mathcal{R}(\widehat{\mathbf{w}}_n)$$

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

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

SQUEAK

Lemma 1. *Assume that the dictionary $\mathcal{I}_{t-1}$ induces a γ-approx. $\widetilde{\mathbf{K}}_{t-1}$, and let $\overline{\mathbf{S}}_t$ be constructed by adding $\overline{q}$ copies of $(\overline{q})^{-1/2}\mathbf{e}_{t,t}$ to the selection matrix. Then, denoting $\alpha = (1+\varepsilon)/(1-\varepsilon)$, for all i such that $i \in \{\mathcal{I}_{t-1} \cup \{t\}\}$,*

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

is an α-approximation of the RLS $\tau_{t,i}$, that is $\tau_{t,i}(\gamma)/\alpha \leq \widetilde{\tau}_{t,i} \leq \tau_{t,i}(\gamma)$.

SQUEAK
Input: Dataset $\mathcal{D}$, regularization $\gamma, \mu, \overline{q}$
Output: $\widetilde{\mathbf{K}}_n$, $\widetilde{\mathbf{w}}_n$
1: Initialize $\mathcal{I}_0$ as empty, $\widetilde{p}_{1,0} = 1$
2: **for** $t = 1, \ldots, n$ **do**
3: Receive new column $[\overline{\mathbf{k}}_t, k_t]$
4: Compute α-app. RLS $\{\widetilde{\tau}_{t,i} : i \in \mathcal{I}_{t-1} \cup \{t\}\}$, using $\mathcal{I}_{t-1}$, $[\overline{\mathbf{k}}_t, k_t]$, and Eq. 4
5: Set $\widetilde{p}_{t,i} = \max\{\min\{\widetilde{\tau}_{t,i}, \widetilde{p}_{t-1,i}\}, \widetilde{p}_{t-1,i}/2\}$
6: Initialize $\mathcal{I}_t = \emptyset$ (Dict-Update)
7: **for all** $j \in \{1, \ldots, t-1\}$ **do** (Shrink)
8: $Q_{t-1,j} = |\{i = j : i \in \mathcal{I}_{t-1}\}|$
9: **if** $Q_{t-1,j} \neq 0$ **then**
10: $Q_{t,j} \sim \mathcal{B}(\widetilde{p}_{t,j}/\widetilde{p}_{t-1,j}, Q_{t-1,j})$
11: Add $Q_{t,j}$ copies of $(j, \mathbf{k}_{t,j}, \widetilde{p}_{t,j})$ to $\mathcal{I}_t$.
12: **end if**
13: **end for**
14: $Q_{t,t} \sim \mathcal{B}(\widetilde{p}_{t,t}, \overline{q})$ (Expand)
15: Add $Q_{t,t}$ copies of $(t, \mathbf{k}_{t,t}, \widetilde{p}_{t,t})$ to $\mathcal{I}_t$
16: Compute $\widetilde{\mathbf{K}}_t$ using $\mathcal{I}_t$, and $\widetilde{\mathbf{w}}_t$ using $\widetilde{\mathbf{K}}_t$, $\mathbf{y}_t$
17: **end for**

Theorem 1. *Let $\alpha = \left(\frac{1+\varepsilon}{1-\varepsilon}\right)$ and $\gamma > 1$. For any $0 \leq \varepsilon \leq 1$, and $0 \leq \delta \leq 1$, if we run SQUEAK with $\overline{q} = \mathcal{O}(\frac{\alpha}{\varepsilon^2}\log(\frac{n}{\delta}))$, then w.p. $1-\delta$, for all $t \in [n]$*

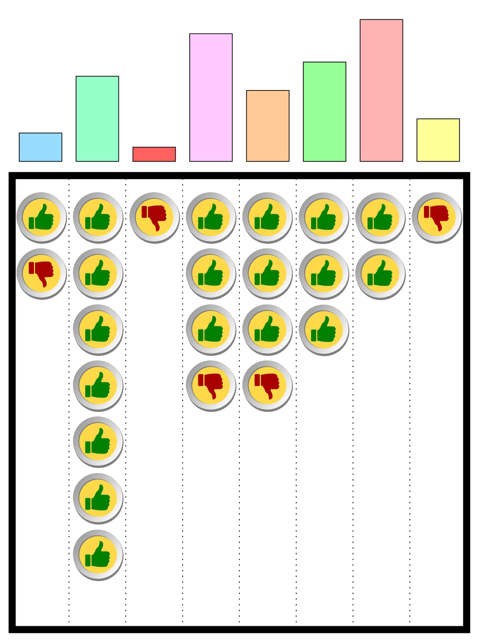
(1) $\widetilde{\mathbf{K}}_t$ *computed with $\mathcal{I}_t$ is a γ-approximation of* $\mathbf{K}_t$.
(2) $|\mathcal{I}_t| = \sum_i Q_{t,i} \leq \mathcal{O}(\overline{q}d_{\text{eff}}(\gamma)_t) \leq \mathcal{O}(\frac{\alpha}{\varepsilon^2}d_{\text{eff}}(\gamma)_n\log(\frac{n}{\delta}))$.
(3) The solution $\widetilde{\mathbf{w}}_t$ satisfies $\mathcal{R}(\widetilde{\mathbf{w}}_t) \leq (1 + \frac{\gamma}{\mu}\frac{1}{1-\varepsilon})\mathcal{R}(\widehat{\mathbf{w}}_t)$.

	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 (C&al'16)	$\frac{\rho^2n^2d_{\text{eff}}{}^2}{\varepsilon^2}$	$\frac{\rho n d_{\text{eff}}}{\varepsilon}$	$(1+2\varepsilon)^2$	Yes
SQUEAK	$\frac{n^2d_{\text{eff}}{}^2}{\varepsilon^2}$	$\frac{d_{\text{eff}}}{\varepsilon^2}$	$(1+2\varepsilon)^2$	Yes

- $\widetilde{\tau}_{t,i} = \mathbf{e}_i^\mathsf{T}\widetilde{\mathbf{K}}_t(\widetilde{\mathbf{K}}_t + \gamma\mathbf{I})^{-1}\mathbf{e}_i$ would fail
- Instead, approximate $\tau_{t,i}$ directly in RKHS $\widetilde{\tau}_{t,i} = \phi(\mathbf{x}_i)^\mathsf{T}(\phi(\mathbf{X}_t)\overline{\mathbf{S}}\overline{\mathbf{S}}^\mathsf{T}\phi(\mathbf{X}_t)^\mathsf{T} + \gamma\mathbf{I})^{-1}\phi(\mathbf{x}_i)$ and then reformulate using kernel trick
- $\widetilde{\tau}_{t,i}$ can be computed in $\mathcal{O}(|\mathcal{I}_t|^2)$ space and $\mathcal{O}(|\mathcal{I}_t|^3)$ time, independent from t.
- $\widetilde{\tau}_{t,i}$ for samples in $\mathcal{I}_t$ can be computed using only samples contained in $\mathcal{I}_t$.
- α trades off accuracy and space/time cost
- The formulation of $\widetilde{\tau}_{t,i}$ is not incremental

Proposition 2. *For any kernel matrix* $\mathbf{K}_{t-1}$ *and its bordering* $\mathbf{K}_t$,

$$\tau_{t,i} \leq \tau_{t-1,i}, \quad d_{\text{eff}}(\gamma)_t \geq d_{\text{eff}}(\gamma)_{t-1}.$$



Pros:
+ Accuracy ***and*** space/time guarantees
+ Unnormalized $\widetilde{p}_{t,i}$, no need for appr. $d_{\text{eff}}(\gamma)_t$
+ In worst case, only $\log(n)$ space overhead
+ Anytime risk guarantees

Cons:
- The time bottleneck is computing intermediate KRR solutions: $\mathcal{O}(t|\mathcal{I}_t|^2)$.
- Still potentially constructs the whole matrix to compute KRR, single pass over matrix but not dataset.

Beyond sequential KRR

What if we run SQUEAK simply to approximate $\mathbf{K}_n$?

- Only need to compute RLS for points in $\mathcal{I}_t$, never recompute after dropping
  - ↳ Never construct the whole $\mathbf{K}_n$, subquadratic runtime ~~$\mathcal{O}(n^2|\mathcal{I}_n|^2)$~~ ⇒ $\mathcal{O}(n|\mathcal{I}_n|^3)$
- Store points directly in the dictionary
  - ↳ $\mathcal{O}(d_{\text{eff}}(\gamma)_n^2 + d_{\text{eff}}(\gamma)_n d)$ space constant in n, single pass over the dataset (streaming)
- Extend Dict-Update (add point to dictionary) to Dict-Merge (add dictionary to dictionary)
  - ↳ Distributed SQUEAK, multiple nodes operate in parallel, without sharing memory recursively merge result to build final dictionary, $\mathcal{O}(\log(n)|\mathcal{I}_n|^3)$ time, $\mathcal{O}(n|\mathcal{I}_n|^3)$ work
- RLS sampling preserves well the projection on $\mathbf{K}_n$'s range

$$\mathbf{P} = \mathbf{K}_n^{1/2}(\mathbf{K}_n + \gamma\mathbf{I})^{-1}\mathbf{K}_n^{1/2} = \phi(\mathbf{X}_n)^\mathsf{T}(\phi(\mathbf{X}_n)\phi(\mathbf{X}_n)^\mathsf{T} + \gamma\mathbf{I})^{-1}\phi(\mathbf{X}_n)$$

  - ↳ SQUEAK provides strong guarantees for many Kernel problems (random/fixed design KRR, Kernel PCA, Kernel k-means)