Randomized Nyström Preconditioning

Zachary Frangella
Cornell

ORIE 7391

February 2021

Quiz

1. Nyström sketch-and-solve is a great method for accurately solving large linear systems.
    - A. True
    - B. False
2. Conjugate gradient works best when the matrix A has a nearly flat spectrum. That is, $\Lambda \approx aI$ for some $a > 0$, where Λ is a diagonal matrix containing the eigenvalues of A and I is the identity matrix.
    - A. True
    - B. False

Introduction

- The focus of this presentation is on solving $(A + \mu I)x = b$, where $A \in \mathbb{R}^{n \times n}$ is symmetric positive definite (spd) and $\mu \geq 0$.
- Symmetric positive linear systems arise in many important applications including: Gaussian processes, kernel ridge regression, optimization etc.
- For many problems, our ability to apply a certain algorithms boils down to whether or not we can solve a symmetric positive definite linear system in a reasonable amount of time.

What's the problem?

- Contemporary problems in machine learning optimization are very large and classical methods for solving linear systems do not scale well.
- Indeed, Cholesky factorization the standard method for solving spd linear systems costs a prohibitive $O(n^3)$.
- Similarly for least-squares, a QR factorization costs $O(Nd^2)$ for an $N \times d$ design matrix X.

Whats the problem? contd.

- Clearly, direct methods such as QR and Cholesky are out.
- This forces to look to iterative methods like Conjugate Gradient (CG) for solving large spd systems.
- CG is much cheaper than QR or Cholesky. It costs only $O(T_{\mathrm{mv}})$ per-iteration.
- This seems good! Maybe things aren't so bad after all.

The problem with CG

- Unfortunately, the situation is not that simple, as CG's convergence is dictated by the condition number

$$\kappa_2(A_\mu) = \frac{\lambda_1 + \mu}{\lambda_n + \mu},$$

  where λ_1 and λ_n are the largest and smallest eigenvalue of A respectively.
- In general, to obtain an ϵ-accurate solution we need to perform $O\left(\sqrt{\kappa_2(A_\mu)}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations of CG.
- Most problems that arising in applications are ill-conditioned. Consequently, CG often converges far too slowly to be practical.

Preconditioning

- The last slide was disappointing, CG seemed liked a perfectly good method before ill-conditioning ruined it.
- Fortunately, there is way to fix it, the solution is known as preconditioning!
- Preconditioning involves solving the equivalent problem,

$$P^{-1/2}A_{\mu}P^{-1/2}z = P^{-1/2}b.$$

- Here P is chosen such that it makes $\kappa_2(P^{-1/2}A_{\mu}P^{-1/2})$ as close to unity is possible, while also being cheap to invert.

Preconditioning contd.

- Solving the preconditioned linear system with CG is referred to as preconditioned Conjugate Gradients (PCG).
- Finding a good preconditioner is often difficult, and is usually problem dependent.
- The main goal of this presentation is to introduce a new preconditioner that is useful for a broad class of problems in machine learning and optimization.

Randomized Nyström Preconditioning

- The new preconditioning approach is called randomized Nyström preconditioning [2]. The technique is based on the randomized Nyström approximation from randomized linear algebra.
- The preconditioner is especially useful in the case of regularized linear systems, where rapid convergence of PCG can be ensured provided we construct the preconditioner appropriately.
- We refer to PCG with a randomized Nyström preconditioner as *Nyström PCG*.

Randomized Nyström Preconditioning contd.

- The key quantity for randomized Nyström preconditioning is the *effective dimension*.
- Given $A \in \mathbb{S}_n^+(\mathbb{R})$ and $\mu \geq 0$, the *effective dimension* of $(A + \mu I)$ is defined by

$$d_{\mathrm{eff}}(\mu) = \mathrm{tr}(A(A+\mu)^{-1}) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \mu}. \tag{1}$$

- Intuitively, $d_{\mathrm{eff}}(\mu)$ is a smoothed count of the eigenvalues larger than μ. It captures the fact that regularization mollifies the eigenvalues smaller than μ.

Preview Nyström PCG vs. CG

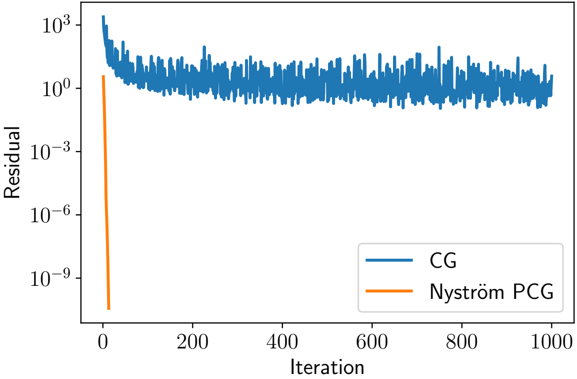

Figure 1: **Ridge regression: CG versus Nyström PCG.** For the shuttle-rf data set, Nyström PCG converges to machine precision in 13 iterations while CG stalls.

Prior work in preconditioning

- One classical technique for constructing a preconditioner is the incomplete Cholesky factorization.
- Incomplete Cholesky however requires structured sparsity patterns to work well, in most ML applications such structure is lacking.
- In the case of least squares A may be very sparse, but $A^T A$ is dense so that incomplete Cholesky is inapplicable.

Prior work in preconditioning contd

- A much more successful approach to preconditioning for ML problems has a arisen in recent years, and is known sketch-and-precondition [5, 1, 4, 3].
- Sketch-and-precondition is designed for highly overdetermined least-squares problems ($X \in \mathbb{R}^{N \times d}$, $N \gg d$).
- Sketch-and-precondition generates a random matrix $S \in \mathbb{R}^{\ell \times n}$ and computes a randomized *sketch* SX, it then performs a pivoted-QR factorization, $SX = QR$ and uses R^{-1} as a preconditioner.

Prior work in preconditioning contd

- Sketch-and-precondition works extremely well in practice but suffers from a severe limitation.
- It requires a sketch size of $\ell = \Omega(d)$, which leads to a cost of $O(T_{\mathrm{mv}}d + d^3)$.
- The $O(d^3)$ cost comes from the pivoted-QR factorization, and limits the applicability of sketch-and-precondition to settings where d is modest in size.

Nyström PCG vs. sketch-and-precondition

- We shall see that Nyström PCG does not require a sketch size of $\ell = \Omega(d)$ to be helpful.
- Nyström PCG applies to square-ish matrices, sketch-and-precondition does not.
- For regularized problems Nyström PCG only requires $\ell = O(d_{\mathrm{eff}}(\mu))$ to ensure fast convergence.

Nyström approximation

- Given $A \in \mathbb{S}_n^+(\mathbb{R})$, a natural way to construct a low-rank approximation is via the Nyström method.
- Let $X \in \mathbb{R}^{n\times\ell}$ be an arbitrary test matrix where $1 \leq \ell \leq n$. The *Nyström approximation* A with respect to X is defined by

$$A\langle X\rangle = (AX)(X^T AX)^{\dagger}(AX)^T. \tag{2}$$

- The Nyström approximation enjoys the following properties.
  - $A\langle X\rangle \in \mathbb{S}_n^+(\mathbb{R})$ and has rank at most ℓ.
  - $A\langle X\rangle \preceq A$, here $\preceq$ denotes the Loewner ordering on $\mathbb{S}_n^+(\mathbb{R})$.
  - Let $\lambda_1 \geq \cdots \geq \lambda_n$ be the eigenvalues of A and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ be the eigenvalues of $A\langle X\rangle$. Then

$$\hat{\lambda}_j \leq \lambda_j \text{ for all } j.$$

Randomized Nyström approximation

- A natural question is how to choose X so that $A\langle X\rangle$ is good low-rank approximation to A?
- It turns out an effective strategy is to take the test matrix to be a Gaussian random matrix Ω [6, 3], which lead us to the *randomized Nyström approximation*.
- The procedure behind the randomized Nyström approximation is as follows.
    - Generate a Gaussian random matrix $\Omega \in \mathbb{R}^{n\times \ell}$.
    - Compute the sketch $Y = A\Omega \in \mathbb{R}^{n\times \ell}$.
    - Form the Nyström approximation

$$\hat{A} = A\langle \Omega\rangle = Y(\Omega^T Y)^\dagger Y^T.$$

Randomized Nyström approximation contd.

- The randomized Nyström approximation requires careful numerical implementation. For the stable implementation see Martinsson & Tropp (2020).
- In the stable implementation $\hat{A}_{\mathrm{nys}}$ is returned in the form of an eigendecomposition,

  $$\hat{A}_{\mathrm{nys}} = U\hat{\Lambda}U^{T}.$$

  where $U \in \mathbb{R}^{n\times\ell}$ and $\hat{\Lambda} \in \mathbb{R}^{\ell\times\ell}$.

Complexity of randomized Nyström approximation

- The computational complexity of $\hat{A}_{\mathrm{nys}}$ when Ω is a Gaussian random matrix is,

$$O(T_{\mathrm{mv}}\ell + n\ell^2).$$

- The storage cost in this setting is $O(n\ell)$.
- A useful property of the randomized Nyström approximation is that it only requires matvecs with A to construct. Thus, $\hat{A}_{\mathrm{nys}}$ can still be constructed when we only have black-box access to A.

A-priori guarantees for randomized Nyström approximation

- The randomized Nyström approximation admits a strong theoretical guarantee on how well it approximates A.

Theorem (Randomized Nyström approximation simplified Gaussian analysis)

Let $A \in \mathbb{S}_n^+(\mathbb{R})$ *with eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_n$*. Let* $p \geq 2$ *and set* $\ell = 2p - 1$*. Draw a Gaussian random test matrix* $\Omega \in \mathbb{R}^{n \times \ell}$*. Then* $\hat{A}_{\text{nys}}$ *satisfies*

$$\mathbb{E}\|A - \hat{A}_{\text{nys}}\| \leq \left(3 + \frac{4e^2}{p}\text{sr}_p(A)\right)\lambda_p,$$

where $\text{sr}_p(A) = \lambda_p^{-1}\sum_{j=p}^{n}\lambda_j$ *is the p-stable rank of A.*

Nyström sketch-and-solve (motivation)

- Suppose we want to solve regularized linear system $(A+\mu I)x = b$.
- Further assume we are in a situation such that $\|E\| = \|A - \hat{A}_{\mathrm{nys}}\|$ is very small and $\ell \ll n$, so that A is well-approximated by $\hat{A}_{\mathrm{nys}}$.
- Then it seems very reasonable to replace A by $\hat{A}_{\mathrm{nys}}$ and instead solve the linear system $(\hat{A}_{\mathrm{nys}} + \mu I)\hat{x} = b$. Furthermore, this system may be solved cheaply, thanks to the following lemma.

Lemma

Let $\hat{A}_{nys} = U\hat{\Lambda}U^T$ be a randomized Nyström approximation of A. Then

$$(\hat{A}_{nys} + \mu I)^{-1} = U(\hat{\Lambda} + \mu I)^{-1}U^T + \frac{1}{\mu}(I - UU^T). \tag{3}$$

Thus, $(\hat{A}_{nys} + \mu I)^{-1}$ may be applied in $O(n\ell)$ time.

Nyström sketch-and-solve

- We refer to the approach on the previous slide as *Nyström sketch-and-solve*, as it closely parallels the classical sketch-and-solve paradigm from randomized linear algebra.
- Nyström sketch-and-solve has a long history in machine learning, where it has been used to speed up kernel ridge-regression.
- Once $\hat{A}_{\mathrm{nys}}$ has been computed, the cost of applying the approximate regularized inverse is $O(n\ell)$.
- The total cost of Nyström sketch-and-solve is is $O(T_{\mathrm{mv}}\ell + n\ell^2)$.

The problem

- Unfortunately, this intuitive and appealing idea has a problem.
- The idea breakdowns when μ becomes small.
- In order to obtain an accurate solution, Nyström sketch-and-solve may require a vacuous rank, $\ell \sim n$.
- As small values of μ frequently arise in practice, this severely restricts Nyström sketch-and-solve's applicability.

Analysis of Nyström sketch-and-solve

- The poor behavior of Nyström sketch-and-solve when μ is small, is a consequence of the following theorem.

Theorem

Let $\hat{A}_{\text{nys}}$ be a randomized Nyström approximation of A constructed from a Gaussian test matrix and let $\hat{x} = (\hat{A}_{\text{nys}} + \mu I)^{-1} b$. Suppose $(A + \mu I)x_\star = b$ and $\hat{A}_{\text{nys}}$ has rank $\ell = 2\lceil 1.5 d_{\text{eff}}(\epsilon\mu)\rceil + 1$ then

$$\mathbb{E}\left[\frac{\|\hat{x} - x_\star\|_2}{\|x_\star\|_2}\right] \leq 26\epsilon. \tag{4}$$

Randomized Nyström preconditioning: motivation

- We have established that $\hat{A}_{\mathrm{nys}}$ is good rank-ℓ approximation to A, it captures the dominant eigenspace of A very well. However, it is not the best rank-ℓ approximation to A.
- Before considering a preconditioner based off of $\hat{A}_{\mathrm{nys}}$, it is instructive to see how well we can do with a preconditioner constructed from the best rank-ℓ approximation.

$$\lfloor A \rfloor_\ell = V_\ell \Lambda_\ell V_\ell^T.$$

- Thus the question becomes: "How well can we reduce the condition number with $\lfloor A \rfloor_\ell$?"

Motivation contd.

- The question on the preceding slide is answered by the following lemma.

Lemma

Let $\mathcal{P} = \{P : P = V_\ell M V_\ell^T + \beta(I - V_\ell V_\ell^T) \quad \textit{where } \beta > 0 \quad \textit{and}$ $M \in \mathbb{S}_\ell^+(\mathbb{R})\}$*. Then for any symmetric psd matrix* A *and* $\mu \geq 0$ *the following holds:*

1.
$$\min_{P \in \mathcal{P}} \kappa_2(P^{-1/2} A_\mu P^{-1/2}) = \frac{\lambda_{\ell+1} + \mu}{\lambda_n + \mu}$$

2.
$$P_\star = \operatorname*{argmin}_{P \in \mathcal{P}} \kappa_2(P^{-1/2} A_\mu P^{-1/2}),$$

where $P_\star = \frac{1}{\lambda_{\ell+1}+\mu} V_\ell \Lambda_\ell V_\ell^T + (I - V_\ell V_\ell^T)$.

Randomized Nyström preconditioning: the preconditioner

- Based on the preceding analysis we propose the following preconditioner for CG based using the randomized Nystöm approximation $\hat{A}_{\text{nys}} = U\hat{\Lambda}U^T$,

$$P = \frac{1}{\hat{\lambda}_\ell + \mu} U(\hat{\Lambda} + \mu I)U^T + (I - UU^T). \tag{5}$$

- The preconditioner requires $O(n\ell)$ storage and may be applied in $O(n\ell)$ time as it admits the explicit inverse,

$$P^{-1} = (\hat{\lambda}_\ell + \mu)U(\hat{\Lambda} + \mu I)^{-1}U^T + (I - UU^T). \tag{6}$$

We refer to P as the *Nyström preconditioner*.

Randomized Nyström preconditioning: controlling the condition number

- We have the following result which shows that the Nyström preconditioner behaves almost like the one constructed from $\lfloor A \rfloor_\ell$.

Theorem (Nyström condition number bound)

Let P be the Nyström preconditioner and $\mu \geq 0$ the ℓ^2-regularization parameter. Let $E = A - \hat{A}_{\text{nys}}$ be the Nyström approximation error. Then the condition number of the preconditioned matrix $P^{-1/2} A_\mu P^{-1/2}$ satisfies

$$\kappa_2(P^{-1/2} A_\mu P^{-1/2}) \leq \min\left\{ \frac{\hat{\lambda}_\ell + \mu + \|E\|}{\mu},\ 1 + \frac{\|E\|}{\hat{\lambda}_\ell + \mu} + \frac{\hat{\lambda}_\ell + \mu + \|E\|}{\lambda_n + \mu} \right\}.$$

An observation

- In the regularized case the preceding theorem implies,

$$\kappa_2(P^{-1/2}A_\mu P^{-1/2}) \leq \frac{\hat{\lambda}_\ell + \mu + \|E\|}{\mu}.$$

- We see that if $\hat{\lambda}_\ell \leq \mu$ and $\|E\| \leq \mu$ then,

$$\kappa_2(P^{-1/2}A_\mu P^{-1/2}) \leq 3.$$

- Consequently, PCG will converge extremely rapidly. This motivates determining the value of ℓ such that $\|E\| \lesssim \mu$. This lead us to back to the *effective dimension*.

The effective dimension

Recall, that the effective dimension is given by

$$d_{\mathrm{eff}}(\mu) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \mu}.$$

We may interpret $d_{\mathrm{eff}}(\mu)$ as a smooth count of the eigenvalues larger than μ.

Given the above interpretation, we might expect that the preconditioned system will be well-conditioned if we choose a target rank of $\ell \sim d_{\mathrm{eff}}$.

Controlling the condition number when $\ell \sim d_{\mathrm{eff}}$

- The preceding intuition is correct and leads to the following theorem.

Theorem

Suppose we construct the Nyström preconditioner P in from a Gaussian test matrix with rank $\ell = 2\lceil 1.5 d_{\mathit{eff}}(\mu)\rceil + 1$. Then the condition number of the preconditioned linear system $P^{-1/2}A_\mu P^{-1/2}$ satisfies

$$\mathbb{E}[\kappa_2(P^{-1/2}A_\mu P^{-1/2})] < 28.$$

Fast convergence of PCG when $\ell \sim d_{\mathrm{eff}}$

- An immediate corollary of the preceding theorem is that PCG converges fast provided the conditions of the theorem are satisfied.

Corollary

Instantiate P as in the theorem on the preceding slide and condition on the event

$$\left\{\kappa_2\left(P^{-1/2}A_\mu P^{-1/2}\right) < 28\right\},$$

Let $M = P^{-1/2}A_\mu P^{-1/2}$. Then starting with initial iterate $x_0 = 0$, the relative error $\delta_t := \|x_t - x_\star\|_M / \|x_\star\|_M$ of each iterate x_t satisfies

$$\delta_t < 2\,(0.69)^t\,.$$

Hence after $t = \lceil 2.7 \log(\frac{2}{\epsilon})\rceil$ iterations of PCG, we have relative error $\delta_t < \epsilon$.

Choosing ℓ in practice

The theory presented shows that when $\ell = O(d_{\text{eff}}(\mu))$, Nyström PCG converges rapidly. Unfortunately, we almost never know the value $d_{\text{eff}}(\mu)$ a-priori, so it is crucial that we have a practical way to choose ℓ. We give three effective strategies for choosing ℓ below.

- Choose ℓ as large as you can afford.
- Adaptive rank selection based on a-posteriori error estimation (estimates of $\|E\|$).
- Adaptive rank selection based the ration $\hat{\lambda}_\ell/\mu$.

Aside from the initial guess supplied by the user, the last two strategies lead to automated rank selection.

Numerical experiments

- To see how the preconditioner behaves in practice, we perform some experiments on random features regression and kernel ridge regression. The datasets and corresponding Kernel hyperparameters are given in the table below. We used the Gaussian kernel for all datasets.

Dataset	n	d	n_{classes}	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	**PCG tolerance**
Higgs-rf	800,000	10,000	2	1e-4	5	1e-10
YearMSD-rf	463,715	15,000	NA	1e-5	8	1e-10
EMNIST-Balanced	105,280	784	47	1e-6	8	1e-3
Santander	160,000	200	2	1e-6	7	1e-3

Table 1: Datasets: statistics and parameters.

Numerical experiments contd.

- For the random features regression problems we compare to the sketch-and-precondition method of Rokhlin and Tygert (R&T) and the Adapative iterative Hessian sketch (AdaIHS) due to Lacotte and Pilanci.
- For the kernel ridge regression experiments we compare to random features preconditioning. The random features preconditioner also admits fast convergence guarantees in terms of $d_{\mathrm{eff}}(\mu)$.
- For the random features regression problems we ran the adaptive version of our algorithm based on a posteriori error estimation.
- For the kernel experiments we simply choose a sketch size of $\ell = 1,000$. for both Nyström and Random features PCG.

Experimental results

Dataset	Method	Number of iterations	Total runtime (s)
Higgs-rf	AdaIHS	55	1,052.7
	R&T	53	607.4
	Adaptive Nyström	28	91.26
YearMSD-rf	AdaIHS	44	1,327.3
	R&T	49	766.5
	Adaptive Nyström	22	209.7
EMNIST	Random features PCG	154	635.2
	Nyström	32	268.4
Santander	Random features PCG	160	810.4
	Nyström	31	164.8

Table 2: The proposed Nyström PCG algorithms outperform all competing methods in both time and iteration complexity.

Summary and conclusion

- Randomized Nyström preconditioning yields a preconditioner that is feasible to construct and behaves like the optimal preconditioner constructed from the best ℓ-approximation to a matrix.
- It can significantly reduce the condition number when the matrix exhibits spectral decay.
- In the case of regularization rapid convergence of PCG can be ensured by choosing a sketch size proportional to the effective dimension.