Matrix-free Stochastic Calculation of Operator Norms without Using Adjoints

on the way to compute the adjoint mismatch

Jonas Bresch

Coauthors: D. A. Lorenz, F. Schneppe, M. Winkler

Table of Contents

Motivation I

- Operator norms are needed in many places, e.g. optimization:
  - Lipschitz constant of the objective $\|Ax - y\|^2$,
  - → gradient descent or the forward-backward method need a stepsize $0 < \tau < 2/\|A\|^2$ for convergence[1],
  - Chambolle-Pock method[2] minimizing $f(x) + g(Ax)$ for convex, closed and proper functions f and g,
  - → needs stepsize σ, τ that fulfill $\sigma\tau \leq 1/\|A\|^2$[3].
- Solvers: power method, requires A^*, if A is not self-adjoint,

→ in discretized setting for $\|A\|$: the adjoint of the discretized one is not the actual adjoint.

---

[1] Ernest K Ryu and Wotao Yin. *Large-scale convex optimization: algorithms & analyses via monotone operators.* Cambridge University Press, 2022.

[2] Antonin Chambolle and Thomas Pock. “A first-order primal-dual algorithm for convex problems with applications to imaging”. In: *Journal of mathematical imaging and vision* 40 (2011), pp. 120–145.

[3] Kristian Bredies et al. “Degenerate preconditioned proximal point algorithms”. In: *SIAM Journal on Optimization* 32.3 (2022), pp. 2376–2401.

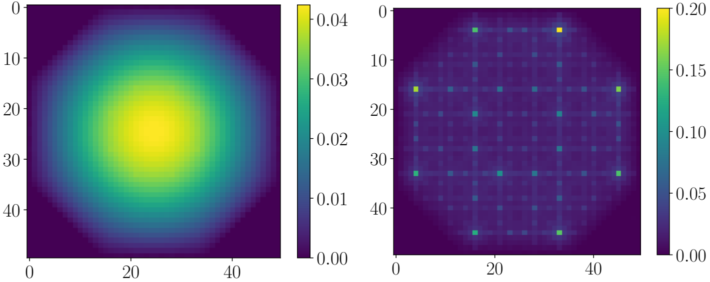
- Rotations of a function $f : \mathbb{D} := \{x \in \mathbb{R}^2 : \|x\| \leq 1\} \to \mathbb{R}$ in $\mathbb{R}^2$,

→ given by $A_\alpha f(x) = f(R_\alpha x)$, where

$$R_\alpha := \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}.$$

- It holds: $A_\alpha^* = A_{-\alpha}$ and $A_\alpha^{-1} = A_{-\alpha}$, such that $\|A_\alpha\| = 1$ in the continues setting.

n	α	$\sqrt{\|\mathbf{A}_{-\alpha}\mathbf{A}_\alpha\|}$	$\|\mathbf{A}_\alpha\|$
25	10	0.99999	1.00463
25	30	1.00002	1.05930
25	45	0.99996	1.17387
50	10	1.00080	0.99804
50	30	0.99998	1.05314
50	45	1.00000	1.16840



Setting

- Operator norm of a linear map between finite dimensional Hilbert spaces, i.e. $\{m, d\}$.
- Unter certain assumption:
  - Evaluations of the linear map $v \mapsto Av$ are available (black-box).
  - Restrictive storage assumptions, i.e. $\mathcal{O}(\max(m, d))$.
  - No access to the adjoint operator A^*.
  - Approximate the value $\|A\|$ to any accuracy.
  - Compute a corresponding maximal singular vector on the fly.

Overview

- We propose:
  - Stochastic algorithm that produces a sequence that converges to the operator norm
  - Uses evaluations of the operator A, and calculation of norms and inner products.
  - For any linear map between finite Hilbert spaces,
  - → do not need any assumptions on the dimension of the spaces or on the structure of A[4].
  - → does not based on the Schatten-p norm[5], for which we need $p \to \infty$, or for computation p sufficiently large enough.
  - → depends on the size of the sketch, does not fulfill out storage assumption.
  - Produces an approximation of a right singular vector for the largest singular value,
  - → extended to compute further right singular vectors.
  - Convergence result for the corresponding eigenvalue and eigenvector equation,
  - **Curiously:** allows us to check if A is orthogonal in the sense that $A^*A = cI$.

[4] Weihao Kong and Gregory Valiant. "Spectrum estimation from samples". In: *Ann. Statist.* 45.5 (2017), pp. 2218–2247. DOI: doi.org/10.1214/16-AOS1525.

[5] Yi Li, Huy L Nguyen, and David P Woodruff. "On sketching matrix norms and the top singular vector". In: *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2014, pp. 1562–1581.

- Random search method,
- with exact, i.e. optimal, step size:
  1) Initialize some $v^0 \in \mathbb{R}^d$ with $\|v^0\| = 1$ and set $k = 0$.
  2) Generate $y^k \sim \mathcal{N}(0, I_d)$.
  3) Project y^k onto the space orthogonal to v_k and normalize it

$$x^k = \frac{y^k - \langle y^k, v^k \rangle v^k}{\|y^k - \langle y^k, v^k \rangle v^k\|}$$

     to get a random normalized search direction perpendicular to v^k.
  4) Calculate a step size τ_k by

$$\tau_k \in \arg\max_{\tau} \frac{\|Av^k + \tau A x^k\|^2}{\|v^k + \tau x^k\|^2}.$$

  5) Update $v^{k+1} = \frac{v^k + \tau_k x^k}{\|v^k + \tau_k x^k\|}$ and go to step 2.

→ **operates fairly simple** if (4) has an exact, and simple to compute, solution.

Table of Contents

First remarks

- By design we always have that $\|v^k\| = \|x^k\| = 1$ and $\langle v^k, x^k \rangle = 0$.
- The random vector x^k is normally distributed on the tangent space on the unit sphere $\mathbb{S}^{d-1}$ at v^k.
- By Rayleigh's principle: critical points of the problem $\max_{\|v\|=1} \|Av\|$ are exactly the right singular vectors of A,
- The intersection of the eigen space of the largest (or smallest) singular value gives exactly the global maxima (or minima).
- All other eigen spaces lead to saddle points.

→ **Crucially:** We show that the method will (a.s.) avoid all saddle points and will converge (a.s.) to the global maximum, i.e. the operator norm.

Table of Contents

Rewriting

- Wirst observations hold by construction

$$\|v^k + \tau x^k\|^2 = \|v^k\|^2 + 2\tau \left\langle v^k, x^k \right\rangle + \tau^2 \|x^k\|^2 = 1 + \tau^2.$$

→ The objective in the maximization problem reads as

$$h_k(\tau) := \frac{\|Av^k + \tau Ax^k\|^2}{1 + \tau^2},$$

→ and derivative

$$h_k'(\tau) = 2\frac{a_k + b_k\tau - a_k\tau^2}{(1 + \tau^2)^2},$$

→ where

$$a_k := \left\langle Av^k, Ax^k \right\rangle \quad \text{and} \quad b_k := \|Ax^k\|^2 - \|Av^k\|^2.$$

→ **All parameters** are available under our certain restrictions.

Solving step size problem

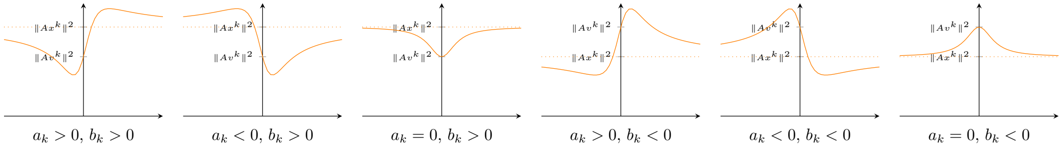
Proposition 1 (Closed form—stepsize)

If $a_k \neq 0$, then the maximizer of h_k (the optimal step size from v^k in direction x^k) is unique and given by

$$\tau_k = \operatorname{sign}(a_k)\left(\frac{b_k}{2|a_k|} + \sqrt{\frac{b_k^2}{4a_k^2} + 1}\right).$$

If $a_k = 0$, then we have:

$$\{\tau_k = 0 : b_k < 0, \qquad \tau_k \to \pm\infty : b_k > 0, \qquad \textit{entire}\, \mathbb{R} : b_k = 0$$



Discuss the cases for a_k

- As long as $a_k \neq 0$, the algorithm leads to a strictly increasing sequence $\|Av^k\|$.
- Convergence follows, since the sequence is bounded by $\|A\|$.
- Our method is a stochastic projected gradient method:
  - $\tau_k > 0$ if and only if $a_k = \langle Av^k, Ax^k \rangle = \langle A^* A v^k, x^k \rangle > 0$,
  - hence, we move in direction $\operatorname{sign}(a_k) x_k$,
  - direction is uniformly distributed on a half-sphere in the space orthogonal to v^k,
  - by symmetry: $\mathbb{E}(\operatorname{sign}(a_k) x^k)$ is a positive multiple of the projection of $A^* A v^k$ on the space orthogonal to v^k,
  - the random search directions fulfill $\mathbb{E}[a_k x^k \mid v^k] = \frac{2}{d-1} \nabla_{\mathbb{S}^{d-1}} \|A \cdot\|^2 [v^k]$.

Lemma 1

Assume: v^k is not a right singular vector of $A : \mathbb{R}^d \to \mathbb{R}^m$.

$$\Rightarrow a_k = \langle Av^k, Ax^k \rangle \neq 0 \ a.s.$$

Theoretical assumptions on A

Lemma 2

Assume:
- σ^2 *is an eigenvalue of* $A^*A : \mathbb{R}^d \to \mathbb{R}^d$ *with geometric multiplicity* $r < d$ *and*
- $v \notin E = \left\{x \in \mathbb{R}^d \mid A^*Ax = \sigma^2 x\right\}$ *the respective eigen space.*

$\Rightarrow \{x \in \{v\}^\perp \mid v + x \in E\}$ *is an affine subspace of dimension* $r-1$ *or is empty.*

Corollary 3

Assume: all singular values of $A : \mathbb{R}^d \to \mathbb{R}^m$ *have multiplicities less than* $d-1$.

If v^k *is not a right singular vector* $\Rightarrow v^{k+1}$ *is a.s. also not a right singular vector.*

Assumption 1

The linear operator $A : \mathbb{R}^d \to \mathbb{R}^m$ *has no singular value with multiplicity* $d-1$ *or* d.

→ **Consequently:** $a_k \neq 0$ (a.s.) throughout the iteration if v^0 is not a right singular vector of A and the sequence $\|Av^k\|$ is increasing and convergent (a.s.).

Almost sure convergence I

- We summarize some basic properties one can prove: For v^k, x^k and a_k holds
  - $\|Av^{k+1}\|^2 - \|Av^k\|^2 = \tau_k a_k \to 0$ a.s. for $k \to \infty$,
  - $a_k = \langle Av^k, Ax^k \rangle \to 0$ a.s. for $k \to \infty$,
  - $\mathbb{E}\big[a_k^2 \mid v^k\big] = \frac{1}{d-1}\big\|\big(I_d - v^k(v^k)^*\big)A^*Av^k\big\|^2 \to 0$ a.s. for $k \to \infty$.

Proposition 2

There exists a subsequence $(v^{k_j})_{j\in\mathbb{N}}$ and a random variable σ which takes values in the singular values of A a.s. such that

- $\operatorname{dist}(E_\sigma, v^{k_j}) \to 0$ *as* $j \to \infty$ *a.s.,*
- $\|Av^k\| \to \sigma$ *as* $k \to \infty$ *a.s.*

- Consider the set $B \coloneqq \{v \in \mathbb{S}^{d-1} : \|Av\|^2 > \sigma_2^2\}$, it holds
  - If $v^k \in B$, then $v^{k+1} \in B$,
  - If $v^{k_0} \in B$ holds for some k_0, then $\lim_{k\to\infty}\|Av^k\| = \sigma_1 = \|A\|$ a.s.

→ **Technically:** we just need $v^{k_0} \in B$ for the final a.s. convergence!

Almost sure convergence II

- Let $\varepsilon > 0$ and $v, v^* \in \mathbb{S}^{d-1}$ and define

$$D_v := \left\{ x \in \{v\}^\perp \mid \overset{\exists \tau \in \mathbb{R} \quad \text{s.t.}}{\frac{v+\tau x}{\|v+\tau x\|} \in B_\theta(\{\pm v^*\})} \right\},$$

- denote by σ_v the uniform distribution on the unit sphere in $\{v\}^\perp$.
- and $B_\theta(\tilde{v}) := \{v \in \mathbb{S}^{d-1} : \langle v, \tilde{v} \rangle \geq \cos(\theta)\}$.

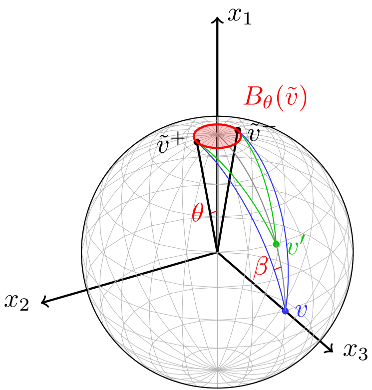
Lemma 4

There exists a positive constant $p_{\varepsilon,d}$ such that

$$\inf_{v \in \mathbb{S}^{d-1}} \sigma_v(D_v) \geq p_{\theta,d}.$$

Theorem 5

- *It holds* $\lim_{k\to\infty} \|Av^k\| = \sigma_1$ *a.s.*
- *If the geometric multiplicity of* σ_1 *is one, then* v^k *converges a.s., too.*



Final remarks

- Stopping criteria: given by $a_k \to 0$,
- combining with $\mathbb{E}[a_k^2 \mid v^k] = \frac{1}{d-1}\|A^*Av^k - \|Av^k\|^2 v^k\|^2$.
- Two singular values: If one singular value has geometrical multiplicity $d-1$, then we either have
  - → $\|Av^0\| > \sigma_2$ if the smallest is of multiplicity $d-1$, and $a_k \neq 0$ by construction,
  - → again $\|Av^0\| > \sigma_2$ if the largest is of multiplicity $d-1$, now, we sample in an $d-1$ dimensional subspace, which is exactly the one corresponding to the largest eigenvector.
- One singular value: we have the following equivalence

$$\mathbb{P}[a_0 = 0] = 1 \quad \Leftrightarrow \quad \mathbb{P}[a_0 = 0] > 0 \quad \Leftrightarrow \quad A^*A = cI_d.$$

Convergence rate

- Study the corresponding singular vector and singular value equation

$$\|A^*Av^k - \|Av^k\|^2 v^k\|^2, \quad k \in \mathbb{N}, \quad \text{where} \quad v^k \text{ is from our method.}$$

- We show a sublinear rates in different settings.

Proposition 3

It holds a.s. $\quad \min_{0\leq k\leq n} a_k^2 \leq \frac{2\|A\|^4}{n+1}$.

Theorem 6

It holds for any $\varepsilon > 0$ *a.s.* $\quad \min_{0\leq k\leq n} \mathbb{P}(\|A^*Av^k - \|Av^k\|^2 v^k\|^2 \geq \varepsilon) \leq \frac{2(d-1)\|A\|^4}{n+1}$.

Table of Contents

Proof of Concept

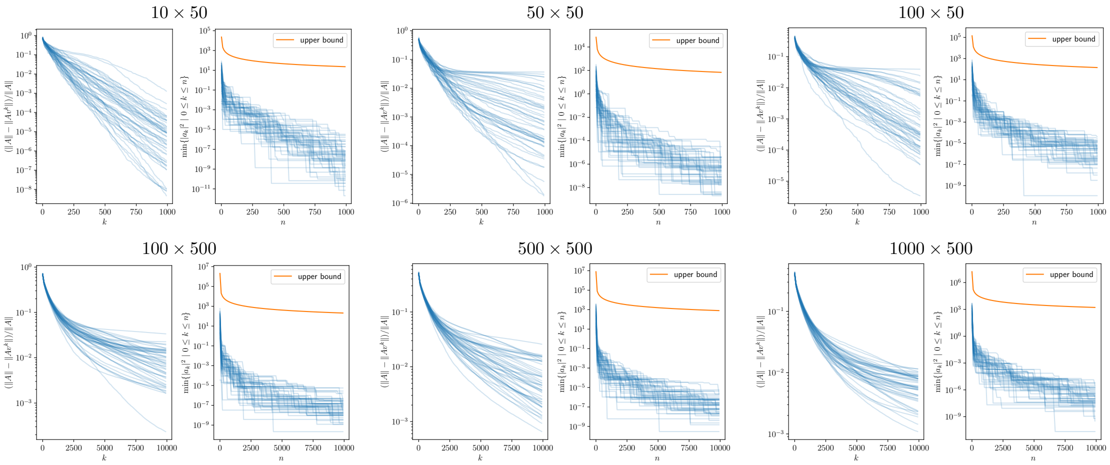
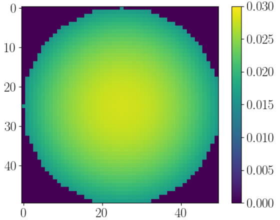
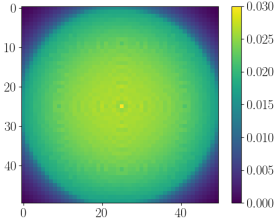


Figure 1: Results for 50 runs of our method for Gaussian matrices of different sizes.

Radon Transformation

- First: from the Python package `scikit-image`[a],
→ provide a *filtered* backprojection in `skimage.transform.iradon`
→ when the filter is set to `None`, approximate adjoint is computed.
- Second: from the `odl` Python package[b],
- **Comparison:**
  1) `skimage`: power method on A^*A: $\|Av^k\| \to 54.84$,
  2) `skimage`: power method on A^*A: $\sqrt{\|A^*Av^k\|} \to 8.22$
  → scaling issue in the adjoint **vs.** our method: 55.86.
  3) `odl`: power method 1.9509808 **vs.** our method: 1.9509840.





[a] Stéfan van der Walt et al. "scikit-image: image processing in Python". In: *PeerJ* 2 (June 2014), e453. DOI: 10.7717/peerj.453.

[b] Frank Natterer and Frank Wübbeling. *Mathematical Methods in Image Reconstruction*. Society for Industrial and Applied Mathematics, 2001. DOI: 10.1137/1.9780898718324.

Table of Contents

Adjoint mismatch calculation

- Consider two linear operators A, V between finite dimensional Hilbert spaces.
- Interested in the problem of the adjoint mismatch $\|A - V\|$.
- **Restrictions:** forwards evaluation for A, i.e. $v \mapsto Av$
- and only backward evaluations for V, i.e. $u \mapsto V^* u$ are available.
- **Notice:** $\|A - V\| = \sup_{u,v} \langle u, (A - V)v \rangle = \sup_{u,v}(\langle V^* u, v\rangle - \langle u, Av\rangle)$.
- **Sketch:** construct u^k, v^k again by orthogonal projections w^k, x^k in $(u^k)^\perp$, i.e. $(v^k)^\perp$ with one global, or two optimal stepsizes for each component.

→ speed up for two directions; one direction less technical in the proof for the a.s. convergence,

→ for two stepsizes, again, the corresponding eigenvalue and eigenvector equation can be studied

$$\min_{0 \le k \le n} \mathbb{P}(\|(A - V)^*(A - V)v^k - a_k^2 v^k\|^2 > \varepsilon) \in \mathcal{O}(n^{-1/2}) \quad \text{a.s.}$$

Thank you for your attention!

References

- Kristian Bredies et al. "Degenerate preconditioned proximal point algorithms". In: *SIAM Journal on Optimization* 32.3 (2022), pp. 2376–2401.
- Antonin Chambolle and Thomas Pock. "A first-order primal-dual algorithm for convex problems with applications to imaging". In: *Journal of mathematical imaging and vision* 40 (2011), pp. 120–145.
- Weihao Kong and Gregory Valiant. "Spectrum estimation from samples". In: *Ann. Statist.* 45.5 (2017), pp. 2218–2247. DOI: doi.org/10.1214/16-AOS1525.
- Yi Li, Huy L Nguyen, and David P Woodruff. "On sketching matrix norms and the top singular vector". In: *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM. 2014, pp. 1562–1581.
- Frank Natterer and Frank Wübbeling. *Mathematical Methods in Image Reconstruction.* Society for Industrial and Applied Mathematics, 2001. DOI: 10.1137/1.9780898718324.
- Ernest K Ryu and Wotao Yin. *Large-scale convex optimization: algorithms & analyses via monotone operators.* Cambridge University Press, 2022.