University of Bremen

Center for Industrial Mathematics

Faculty 03
Mathematics/Computer Science

Optimization without (exact) adjoints

Dirk Lorenz
German-Polish Conference on Optimization

Outline

1. Introduction
2. Minimize $||Av - b||$
3. Minimizing $F(Ax) + G(x)$
4. $||A||$ and $||A - V||$
5. Conclusion and Outlook

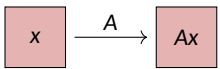
Things we do in optimization

For a linear map A (and a vector b) we…

- → compute least squares solutions to $Ax = b$
- → solve minimization problem of the form

$$\underset{x}{\text{minimize}}\, F(Ax) + G(x)$$

- → compute $\|A\|$



Things we do in optimization

For a linear map A (and a vector b) we…

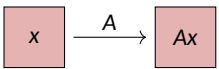
- → compute least squares solutions to $Ax = b$
- → solve minimization problem of the form

$$\underset{x}{\text{minimize}}\, F(Ax) + G(x)$$

- → compute $\|A\|$

but now, consider the situation where

- → $A : X \to Y$ is given as a black box operator: For given $x \in X$ you get $Ax \in Y$.



Things we do in optimization

For a linear map A (and a vector b) we…

- → compute least squares solutions to $Ax = b$
- → solve minimization problem of the form

$$\underset{x}{\text{minimize}}\, F(Ax) + G(x)$$

- → compute $\|A\|$

but now, consider the situation where

- → $A : X \to Y$ is given as a black box operator: For given $x \in X$ you get $Ax \in Y$.
- → Nothing else. (Well, almost…)

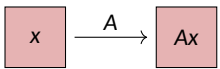
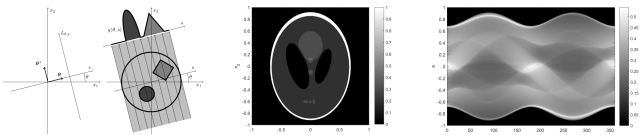


Exhibit A: The forward and backward projection in the Radon transform

→ Basic of many tomography methods: **Radon transform** of $f : \Omega \to \mathbb{R}, \Omega \subset \mathbb{R}^2$

$$\mathcal{R}[f](\theta, s) = \int_{L_{\theta,s}} f \, d\ell, \quad L_{\theta,s} = \left\{ \vec{x} \,\middle|\, \vec{x} \cdot \vec{\theta} = s \right\}, \quad \vec{\theta} = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}, \quad s \geq 0.$$



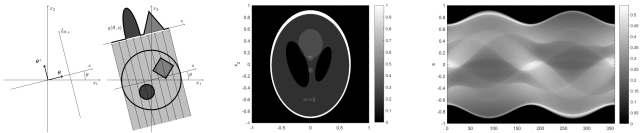
Hansen, Jørgensen Lionheart, Computerized tomography

→ For $\Omega = B_1(0)$: $\mathcal{R} : L^2(\Omega) \to L^2([0, 2\pi[\times[0, 1])]$ linear and bounded
→ Discretizations A: `radon` in MATLAB, Python's skimage, the ASTRA toolbox, gratopy…
→ Adjoint operation: Backprojection V^*
→ Folk wisdom: *If $V^* = A^*$ exactly, only one of them can be fast.* ⇝ People use $V^* \approx A^*$.

Exhibit A: The forward and backward projection in the Radon transform

→ Basic of many tomography methods: **Radon transform** of $f : \Omega \to \mathbb{R}, \Omega \subset \mathbb{R}^2$

$$\mathcal{R}[f](\theta, s) = \int_{L_{\theta,s}} f \, d\ell, \quad L_{\theta,s} = \left\{ \vec{x} \,\middle|\, \vec{x} \cdot \vec{\theta} = s \right\}, \quad \vec{\theta} = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}, \quad s \geq 0.$$



Hansen, Jørgensen Lionheart, Computerized tomography

→ For $\Omega = B_1(0)$: $\mathcal{R} : L^2(\Omega) \to L^2([0, 2\pi[\times[0,1])]$ linear and bounded
→ Discretizations A: `radon` in MATLAB, Python's skimage, the ASTRA toolbox, gratopy…
→ Adjoint operation: Backprojection V^*
→ Folk wisdom: *If $V^* = A^*$ exactly, only one of them can be fast.* ⇝ People use $V^* \approx A^*$.
→ So even in one of the best-understood inverse problems, the adjoint you use is already wrong.

Exhibit B: Rotations

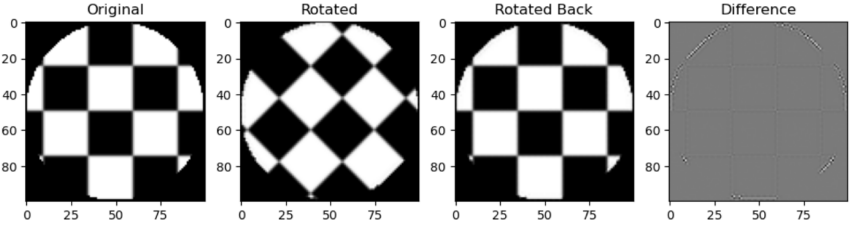
→ Consider $R : L^2(\Omega) \to L^2(\Omega)$,

$$Rf(x) = f(\mathbf{R}_\alpha x)$$

with a rotation matrix $\mathbf{R}_\alpha$.

→ R unitary with $R^* f(x) = R^{-1} f(x) = f(\mathbf{R}_\alpha^{-1} x) = f(\mathbf{R}_{-\alpha} x)$.

→ Discretization uses interpolation. scipy's `rotate` with bicubic interpolation:



relative error 3.8%

→ So numerical errors result in adjoint mismatch naturally.

Can we really...

- $\underset{v}{\text{minimize}} \|Av - b\|^2$
- $\underset{x}{\text{minimize}}\, F(Ax) + G(x)$
- compute $\|A\|$
- compute $\|A - V\|$

when

- $A : X \to Y$ is given as a black box operator: For given $x \in X$ you get $Ax \in Y$,
- $V^* : Y \to X$ is an approximate adjoint?

Let's see how to handle these tasks without using any adjoint!

Outline

1. Introduction
2. Minimize $||Av - b||$
3. Minimizing $F(Ax) + G(x)$
4. $||A||$ and $||A - V||$
5. Conclusion and Outlook

Adjoint free methods for least squares problems

- Things that don't work:
  - Gradient based minimization $\nabla \|Av - b\|^2 = A^T(Av - b)$ needs the transpose…
  - Conjugate gradients (CG) naturally need adjoint
  - GMRES needs $X = Y$ (square matrices)
- Existing methods:
  - "Transpose free quasi minimial residual method" TFQMR [Freund, 1993]
  - "Conjugate gradient squared" CGS [Sonneveld, 1989, Brezinski et al., 1998, Chan et al., 1998]
  - randomized iterative methods [Gower, Richtarik, 2015]

Adjoint sampling

→ Simple trick: If x is a random vector with $\mathbb{E}[xx^T] = I$, then

$$\mathbb{E}[\langle v, Ax\rangle x] = A^* v.$$

Proof: $\mathbb{E}[\langle v, Ax\rangle x] = \mathbb{E}[xx^T A^T v] = A^T v.$ □

Theorem: [L., Schneppe, Tondji, 2023] The iterates of $v^{k+1} = v^k + \tau_k \langle Av^k - b, Ax\rangle x$ converge to a solution of the consistent system $Av = b$ when $\tau_k = \tau$ and $0 < \tau < 2/(c\|A\|^2)$ where $\mathbb{E}[xx^T \|x\|^2] = cI$.

Adjoint sampling is good enough for convergence, second-moment matching is the *only* requirement – tails, boundedness, sparsity of x are free design choices.

Which samplers work?

→ Need $\mathbb{E}[xx^T] = I \in \mathbb{R}^{d\times d}$ and $\mathbb{E}[xx^T\|x\|^2] = cI$ for some c:

- → **Normally distributed**, $x \sim \mathcal{N}(0, I)$: coordinates independent, unit variance $\Rightarrow \mathbb{E}[xx^T] = I$ directly.
  One calculates $\mathbb{E}[xx^T\|x\|^2] = (d+2)I$
- → **Rademacher vectors**, $x_i \in \{-1, 1\}$ i.i.d.: $\mathbb{E}[x_i x_j] = \delta_{ij}$ since $\mathbb{E}[x_i^2] = 1$, $\mathbb{E}[x_i] = 0$.
  Since $\|x\|^2 = d$ deterministically, we have $\mathbb{E}[xx^T\|x\|^2] = dI$
- → **Random coordinate directions**, $x = \sqrt{d}\, e_k$, k uniform on $\{1, \dots, d\}$: $\mathbb{E}[xx^T] = d \cdot \frac{1}{d}\sum_k e_k e_k^T = I$.
  Again $\|x\|^2 = d$ deterministically, we have $\mathbb{E}[xx^T\|x\|^2] = dI$
- → **Uniformly on the sphere**, $x \sim \mathrm{Unif}(\sqrt{d}\,\mathbb{S}^{d-1})$: rotational symmetry forces $\mathbb{E}[xx^T] = cI$; taking the trace gives $c = 1$.
  Again $\|x\|^2 = d$ deterministically, we have $\mathbb{E}[xx^T\|x\|^2] = dI$

Four different distributions, one identity: $\mathbb{E}[\langle v, Ax\rangle x] = A^* v$. Randomness replaces the transpose.

Faster minimization, by exact linesearch

→ Aim is to make $\|Av - b\|^2$ small. If direction x is given, we can minimize $\|A(v + \tau x) - b\|^2$ exactly: $\tau = -\frac{\langle Av - b, Ax\rangle}{\|Ax\|^2}$ (if $\|Ax\| \neq 0$).

Theorem: [L., Schneppe, Tondji 2023] The iterates of $v^{k+1} = v^k - \frac{\langle Av^k - b, Ax\rangle}{c\|A\|^2} x$ fulfill

$$\mathbb{E}(\|Av^{k+1} - b\|^2) \leq (1 - \frac{\sigma_{\min}(A)^2}{c\|A\|^2})\mathbb{E}(\|Av^k - b\|^2)$$

and for $v^{k+1} = v^k - \frac{\langle Av^k - b, Ax\rangle}{\|Ax\|^2} x$ we have

$$\mathbb{E}(\|Av^{k+1} - b\|^2) \leq (1 - \lambda_{\min}(AMA^T))\mathbb{E}(\|Av^k - b\|^2)$$

where $M = \mathbb{E}\left(\frac{xx^T}{\|Ax\|^2}\right)$.

Exact linesearch, no step-size tuning, random directions lead to linearly convergent method.

Numerical results with random matrices

		$\frac{\|Av-b\|}{\|b\|}$	$\|v\|$	time (s)
Rademacher	SGDAS	8.69e-01	1.97e+00	2.42e+00
	RD	9.97e-03	1.74e+01	1.00e+00
Coordinate	SGDAS	9.99e-01	8.54e-03	2.05e+00
	RD	9.56e-03	1.74e+01	9.27e-01
Spherical	SGDAS	9.99e-01	8.52e-03	2.20e+00
	RD	9.98e-03	1.74e+01	8.60e-01
Normal	SGDAS	8.69e-01	1.96e+00	2.09e+00
	RD	9.99e-03	1.74e+01	8.62e-01
TFQMR		3.16e+00	2.78e+02	1.44e+00
CGS		1.96e+01	9.89e+02	2.36e+00

Table: Comparison of SGDAS, RD, TFQMR and CGS. Stopped at relative tolerance of 10^{-2} or after 10000 iterations. Random sparse matrices A with normally distributed entries and random solution vectors $\hat{v}$ with normally distributed entries. Size of A: 1200×300, density of A: 0.1.

CGS and TFQMR fail, random descent works, even fairly fast

Numerical results with matrices from SuiteSparse

Name	m	d	Method	$\frac{\|Av-b\|}{\|b\|}$	time (s)
ash292	292	292	RD Rademacher	1.09e-01	7.12e-01
			RD Normal	1.15e-01	8.40e-01
			TFQMR	2.49e-03	1.09e-01
			CGS	9.50e-03	2.73e-02
ash331	331	104	RD Rademacher	1.00e-02	2.61e-01
			RD Normal	9.79e-03	2.43e-01
			TFQMR	9.92e-01	1.09e+00
			CGS	6.48e+25	1.51e+00
ash608	608	188	RD Rademacher	9.98e-03	9.81e-01
			RD Normal	9.99e-03	1.50e+00
			TFQMR	1.00e+00	3.03e+00
			CGS	1.90e+26	5.40e+00

Table: Comparison of RD, TFQMR, and CGS on least squares problems from the SuiteSparse Matrix collection.

CGS and TFQMR can fail dramatically, random descent always works, no adjoint needed

Recap: adjoint-free least squares

- → No A^T needed: random directions give an unbiased gradient estimator
- → Leads to a *linear* rate in expectation
- → In practice: Exact line search even competitive with (or better than) TFQMR/CGS, and more robust

Outline

1. Introduction
2. Minimize $||Av - b||$
3. Minimizing $F(Ax) + G(x)$
4. $||A||$ and $||A - V||$
5. Conclusion and Outlook

Optimization with adjoint mismatch

- [Zeng, Gullberg 2001], [Elfving, Hansen 2018] and [Dong et al. 2019] for ART methods in tomography
- [L., Schöpfer, Rose 2018] for randomized Kaczmarz
- [Savanier et al. 2021] and [Chouzenoux et al. 2021] for proximal gradient
- [Chouzenoux et al. 2023] for Condad-Vũ, Loris-Verhoeven and Combettes-Pesquet for $\text{minimize}_x \frac{1}{2}\|y - Hx\|^2 + F(Ax) + G(x)$ and mismatch for H
- [Naldi, Schneppe 2025] for primal-dual Douglas-Rachford
- Here: Chambolle-Pock for $\text{minimize}_x F(Ax) + G(x)$ and mismatch in A

Results from [L., Schneppe 2022]

- Let F, G closed, convex, proper on Hilbert spaces and consider $\text{minimize}_x\, F(Ax) + G(x)$
- Proximal operators:

$$\text{prox}_{\tau G}(x) = \underset{y}{\text{argmin}}\, \tfrac{1}{2}\|x - y\|^2 + \tau G(y)$$

- Chambolle-Pock with mismatch in the adjoint

$$x^{i+1} = \text{prox}_{\tau_i G}(x^i - \tau_i V^* y^i)$$
$$y^{i+1} = \text{prox}_{\sigma_i F}(y^i + A(x^{i+1} + \omega_i(x^{i+1} - x^i)))$$

- Error bound: If G γ_G-strongly convex, x^* the minimizer of $F(Ax) + G(x)$, and $(\hat{x}, \hat{y})$ fixed-point of mismatched iteration, then

$$\|x^* - \hat{x}\| \leq \tfrac{1}{\gamma_G}\|(V - A)\hat{y}\|.$$

The reconstruction error is linear in the adjoint mismatch $\|A - V\|$ – small mismatch, small error.

Results from [L., Schneppe 2022]

→ Fixed-points fulfill

$$0 \in \begin{pmatrix} \partial G & V^* \\ -A & \partial F^* \end{pmatrix} \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} =: \mathcal{A}\hat{u}$$

→ $\mathcal{A}$ monotone if F^* γ_{F^*}-strongly convex and $\gamma_{F^*}\gamma_G > 2\|A - V\|$.

→ Under same condition: There exist stepsizes $\tau, \sigma > 0$ and $\omega \in]0,1[$ such that

$$\|x^n - \hat{x}\|^2 + \|y^n - \hat{y}\|^2 = \mathcal{O}(\omega^n).$$

Strong convexity measured by $\gamma_{F^*}\gamma_G$ can absorb fixed amount of adjoint mismatch $\|A - V\|$.

Recap: Chambolle-Pock with adjoint mismatch

- Chambolle-Pock method for minimizing $F(Ax) + G(x)$ not very stable under adjoint mismatch.
- Convergence only proven for strong convexity of G *and* F^* (smooth F)
- Stepsize choice uses bounds on $\|V^*\|$ and $\|A - V\|$.

Outline

1. Introduction
2. Minimize $||Av - b||$
3. Minimizing $F(Ax) + G(x)$
4. $||A||$ and $||A - V||$
5. Conclusion and Outlook

Calculating $||A||$ by maximizing on the sphere

→ Idea: Calculate $||A||^2 = \max_{||v||=1} ||Av||^2$ using an ascent algorithm with random directions.

→ Some tricks that make this work:

- → At some $v^k \in \mathbb{S}^{d-1}$ sample x^k uniformly with $x^k \perp v^k$ and $||x|| = 1$.
- → The linesearch $\underset{\tau}{\text{maximize}} \frac{||Av^k + \tau Ax^k||^2}{||v^k + \tau x^k||^2}$ can be solved exactly:

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

where $a_k = \langle Av^k, Ax^k \rangle$, $b_k = ||Ax^k||^2 - ||Av^k||^2$

Random tangential directions, exact linesearch possible analytically

Theorem: [Bresch, L., Schneppe, Winkler, 2024] If $a_0 \neq 0$, then all $a_k \neq 0$ almost surely and

$$\|Av^k\| \to \|A\| \quad \text{a.s.}$$

If the singular space for $\sigma_1 = \|A\|$ is one-dimensional, then v^k converges almost surely to a corresponding singular vector.
Moreover

$$\min_{0\leq k\leq n} \mathbb{P}\left(\|A^*Av^k - \|Av^k\|^2 v^k\|^2 \geq \varepsilon\right) \leq \frac{2(d-1)\|A\|^4}{(n+1)\varepsilon}.$$

→ For $\|A - V\|$ use $\|A - V\| = \max_{\|u\|=\|v\|=1} \langle u, (A-V)v\rangle = \max_{\|u\|=\|v\|=1} [\langle u, Av\rangle - \langle V^*u, v\rangle]$ and do random ascent in both u and v

No SVD, no adjoint, no assembled matrix – just forward evaluations of A, and $\|A\|$ is still found almost surely.

How large is the norm of a discrete rotation?

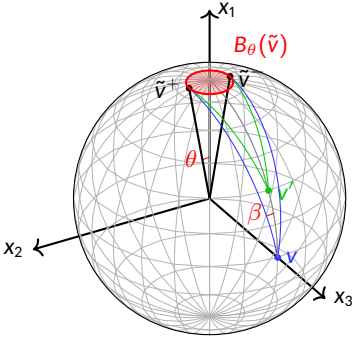
- Come back to rotation operator $Rf(x) = f(\mathbf{R}_\alpha x)$
- $\mathbf{A}_\alpha$ discretized $\mathbf{R}_\alpha$ for $n \times n$ images with bicubic interpolation.
- Compute norm by power iteration for $\mathbf{A}_{-\alpha}\mathbf{A}_\alpha$ vs. adjoint free method:

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

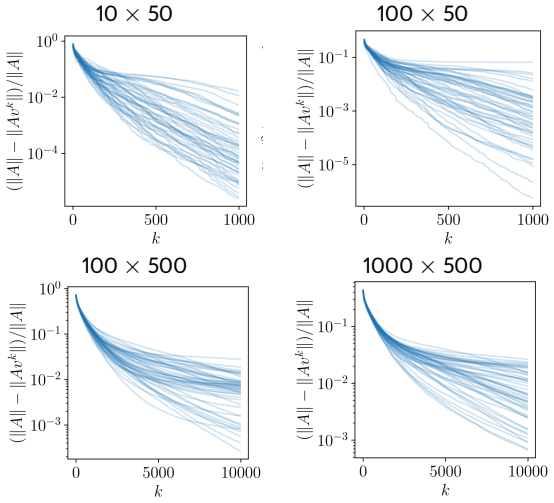
The continuous rotation is unitary ($\|R\| = 1$), but its discretization is not: $\|\mathbf{A}_\alpha\|$ can be considerably larger.

Why does this method converge a.s.?

- The sequence $\|Av^k\|$ is strictly increasing a.s.
- The only cluster points are singular vectors
- Once $\|Av^k\| > \sigma_2$ we know that we converge to σ_1
- The probability that the next iterate is close to $\tilde{v}$ is positive and uniformly bounded from below:



Convergence behavior



Recap: matrix-free norms

- → Random ascent with exact linesearch finds $||A||$ almost surely, without ever forming A or A^T
- → The same machinery computes $||A - V||$ – so the size of the mismatch itself becomes measurable
- → Confirmed numerically, even for a genuinely non-self-adjoint operator (the discrete rotation)

Outline

1. Introduction
2. Minimize $||Av - b||$
3. Minimizing $F(Ax) + G(x)$
4. $||A||$ and $||A - V||$
5. Conclusion and Outlook

Conclusion

- When the exact adjoint is not available it can be replaced by statistical estimates (another way not followed here: autodiff for $(x, y) \mapsto \langle y, Ax \rangle$ w.r.t. x)
- Exact line searches greatly help for least squares problems and norm estimation, making the methods practically relevant.
- Even for the general purpose problem $\text{minimize}_x\, F(Ax) + G(x)$ convergence can be achieve if both G and F^* are strongly convex.
- Meaningful optimization is possible even without adjoint or with only approxmate adjoint information

Outlook

- Nonlinear forward operators $F : X \to Y$: replace the adjoint by directional derivatives – does randomization still avoid the transpose?
- Sharper, non-asymptotic mismatch bounds for Chambolle-Pock beyond the strongly-convex case
- Accelerate norm computation by larger search spaces (produce low-dimensional sketch, do local exact optimization from there on)

Thanks for listening!

1. Introduction
2. Minimize $\|Av - b\|$
3. Minimizing $F(Ax) + G(x)$
4. $\|A\|$ and $\|A - V\|$
5. Conclusion and Outlook


Jonas Bresch, Dirk A Lorenz, Felix Schneppe, and Maximilian Winkler.
Computing adjoint mismatch of linear maps.
Journal of Computational and Applied Mathematics, page 117853, 2026.


Jonas Bresch, Dirk A Lorenz, Felix Schneppe, and Maximilian Winkler.
Matrix-free stochastic calculation of operator norms without using adjoints.
SIAM Journal on Matrix Analysis and Applications, 47(3):1349–1374, 2026.


Dirk A. Lorenz and Felix Schneppe.
Chambolle-Pock's primal-dual method with mismatched adjoint.
Applied Mathematics and Optimization, 87(22), 2023.


Dirk A. Lorenz, Felix Schneppe, and Lionel Tondji.
Linearly convergent adjoint free solution of least squares problems by random descent.
Inverse Problems, 39(12):125019, 2023.