Technische Universität Braunschweig

Random descent for least squares functionals

Dirk Lorenz, joint work with Felix Schneppe and Lionel Tondji, May 22, 2023

Linear least squares

Consider simple, plain least squares

$$\min_{v\in\mathbf{R}^d} \tfrac{1}{2}\|Av-b\|^2, \qquad A\in L(\mathbf{R}^d,\mathbf{R}^m),\ b\in\mathbf{R}^m$$

but under the following assumptions:

Linear least squares

Consider simple, plain least squares

$$\min_{v \in \mathbf{R}^d} \tfrac{1}{2}\|Av - b\|^2, \qquad A \in L(\mathbf{R}^d, \mathbf{R}^m),\ b \in \mathbf{R}^m$$

but under the following assumptions:

1. There is only an implementation available, that produces Ax for any x

Linear least squares

Consider simple, plain least squares

$$\min_{v\in\mathbf{R}^d} \tfrac{1}{2}\|Av - b\|^2, \qquad A \in L(\mathbf{R}^d, \mathbf{R}^m), b \in \mathbf{R}^m$$

but under the following assumptions:

1. There is only an implementation available, that produces Ax for any x
2. There is no implementation of the adjoint/transpose available

Linear least squares

Consider simple, plain least squares

$$\min_{v\in\mathbf{R}^d} \tfrac{1}{2}\|Av - b\|^2, \qquad A \in L(\mathbf{R}^d, \mathbf{R}^m), b \in \mathbf{R}^m$$

but under the following assumptions:

1. There is only an implementation available, that produces Ax for any x
2. There is no implementation of the adjoint/transpose available
3. Automatic differentiation does not work for that implementation (otherwise could get $A^T y$ as derivative of $\langle y, Ax\rangle$ w.r.t. x).

Linear least squares

Consider simple, plain least squares

$$\min_{v\in\mathbf{R}^d} \tfrac{1}{2}\|Av - b\|^2, \qquad A \in L(\mathbf{R}^d, \mathbf{R}^m), b \in \mathbf{R}^m$$

but under the following assumptions:

1. There is only an implementation available, that produces Ax for any x
2. There is no implementation of the adjoint/transpose available
3. Automatic differentiation does not work for that implementation (otherwise could get $A^T y$ as derivative of $\langle y, Ax\rangle$ w.r.t. x).
4. Only a very small number of vectors of size either d or m can be stored (i.e. no way to approximate a larger portion of matrix representation of A)

Linear least squares

Consider simple, plain least squares

$$\min_{v\in\mathbf{R}^d} \tfrac{1}{2}\|Av-b\|^2, \qquad A\in L(\mathbf{R}^d,\mathbf{R}^m),\, b\in\mathbf{R}^m$$

but under the following assumptions:

1. There is only an implementation available, that produces Ax for any x
2. There is no implementation of the adjoint/transpose available
3. Automatic differentiation does not work for that implementation (otherwise could get $A^T y$ as derivative of $\langle y, Ax\rangle$ w.r.t. x).
4. Only a very small number of vectors of size either d or m can be stored (i.e. no way to approximate a larger portion of matrix representation of A)

Linear least squares

Consider simple, plain least squares

$$\min_{v\in\mathbf{R}^d} \tfrac{1}{2}\|Av - b\|^2, \qquad A \in L(\mathbf{R}^d, \mathbf{R}^m),\ b \in \mathbf{R}^m$$

but under the following assumptions:

1. There is only an implementation available, that produces Ax for any x
2. There is no implementation of the adjoint/transpose available
3. Automatic differentiation does not work for that implementation (otherwise could get $A^T y$ as derivative of $\langle y, Ax\rangle$ w.r.t. x).
4. Only a very small number of vectors of size either d or m can be stored (i.e. no way to approximate a larger portion of matrix representation of A)

What can we still do?

Adjoint sampling

Lemma

If $x \in \mathbf{R}^d$ *is a random vector with* $\mathbf{E}(xx^T) = I_d$*, then*

$$\mathbf{E}(\langle Av - b, Ax\rangle x) = A^T(Av - b),$$

i.e. $\langle Av - b, Ax\rangle x$ *is an unbiased estimate for* $\nabla(\frac{1}{2}\|Av - b\|^2)$.

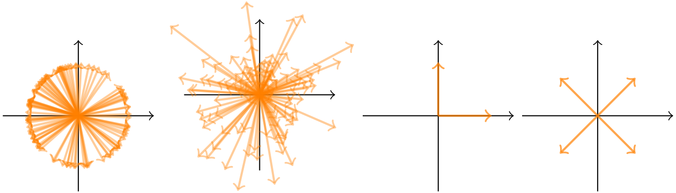
Hence, we can do *stochastic gradient descent:*

$$v^{k+1} = v^k - \tau_k \langle Av^k - b, Ax\rangle x \quad \text{for } x \sim \mathcal{D}$$

Isotropic random vectors

Random vector $x \sim \mathcal{D}$ is *isotropic* if $\mathbf{E}(xx^T) = I_d$

- Random unit vectors: $x \sim \mathrm{Unif}(\sqrt{d}S^{d-1})$
- Standard normal vectors $x \sim \mathcal{N}(0, I_d)$
- Random coordinate vectors $\mathbf{P}(x = \sqrt{d}e_k) = \frac{1}{d}$
- Rademacher vectors $\mathbf{P}(x_k = \pm 1) = \frac{1}{2}$ independently



- Necessarily:
  $\mathbf{E}(\|x\|^2) = \mathbf{E}(x^T x) = \mathbf{E}(\mathrm{trace}(x^T x)) = \mathbf{E}(\mathrm{trace}(xx^T)) = \mathrm{trace}(\mathbf{E}(xx^T)) = \mathrm{trace}(I_d) = d$

Stochastic gradient descent with adjoint sampling (SGDAS)

Random isotropic x ($\mathbf{E}(xx^T) = I_d$) which also fulfills $\mathbf{E}(xx^T\|x\|^2) = cI_d$

Initialize $\nu^0 = 0 \in \mathbf{R}^d$, $k = 0$, $\tau > 0$
while not stopped **do**
obtain random vector $x \in \mathbf{R}^d$
update $\nu^{k+1} = \nu^k - \tau\langle A\nu^k - b, Ax\rangle x$
end while

Theorem

Let $A\hat{\nu} = b$ and $(\nu^k)_k$ generated by SGDAS with $0 < \tau < 2/(c\|A\|^2)$. Then

$$\mathbf{E}(\|\nu^{k+1} - \hat{\nu}\|^2) \leq \lambda^k\|\nu^0 - \hat{\nu}\|^2$$

for $\lambda = \|I - \tau A^TA(2I - \tau cA^TA)\|$. Esp. $\tau = \frac{2}{c}\frac{1}{\lambda_{\max}+\lambda_{\min}}$ gives $\lambda = 1 - \frac{4\kappa(A)}{c(\kappa(A)+1)^2}$

Convergence of residuals

Theorem

Let $(v^k)_k$ be generated by SGDAS with $\tau_k = \tau = \frac{1}{c\|A\|^2}$. Then it holds that

$$\min_{0\leq k\leq N-1} \mathbf{E}(\|A^T(Av^k - b)\|^2) \leq \frac{c\|A\|^2\|b\|^2}{N}.$$

Theorem

With $\beta = 1 - \tau\sigma_{\min}(A)^2(2 - \tau c\|A\|^2)$ it holds that

$$\mathbf{E}(\|Av^{k+1} - b\|^2) \leq \beta^{k+1}\|Av^0 - b\|^2.$$

$\tau = 1/(c\|A\|^2)$ gives $\beta = 1 - \frac{\sigma_{\min}(A)^2}{c\|A\|^2}$

Inconsistent systems

Theorem

Let $A\hat{v} = b$ *and* $(v^k)_k$ *be generated by SGDAS with* $0 < \tau < \frac{2}{c\|A\|^2}$*, and rhs* $\tilde{b} = b + r$ *with* $r = r' + r''$ *with* $r' \in \mathrm{rg}(A)$ *and* $r'' \in \mathrm{rg}(A)^\perp$*. Then*

$$\mathbf{E}(\|v^{k+1} - \hat{v}\|^2) \leq \left(\frac{1+\lambda}{2}\right)^{k+1} \|v^0 - \hat{v}\|^2 + \tau^2 \frac{2\left((1-\lambda)c + 2\|I - \tau c A^T A\|^2\right)}{(1-\lambda)^2} \|A^T r'\|^2$$

Drawbacks of SGDAS:

- Very slow rate (note division by c; holds: $c > d$)
- Stepsize needs knowledge about $\|A\|$ (how to calculate without using A^T?)

$\rightsquigarrow$ Try linesearch

Intermission: Calculating norms without adjoints

1. How to calculate $\|A\|$ given our constraints?
2. Even more difficult: Assume that only routines for $x \mapsto Ax$ and $y \mapsto V^T y$ available. How to calculate $\|A - V\|$?

Intermission: Calculating norms without adjoints

1. How to calculate $\|A\|$ given our constraints?
2. Even more difficult: Assume that only routines for $x \mapsto Ax$ and $y \mapsto V^T y$ available. How to calculate $\|A - V\|$?

- For 1. use stochastic coordinate ascent to solve $\|A\|^2 = \max_{\|\nu\|=1} \|A\nu\|^2$ with adjoint sampling:
$$\nu^{k+1/2} = \nu^k + \tau_k \langle A\nu^k, Ax\rangle x, \quad \nu^{k+1} = \frac{\nu^{k+1/2}}{\|\nu^{k+1/2}\|}$$
- Linesearch more difficult, but possible...

Intermission: Calculating norms without adjoints

1. How to calculate $\|A\|$ given our constraints?
2. Even more difficult: Assume that only routines for $x \mapsto Ax$ and $y \mapsto V^T y$ available. How to calculate $\|A - V\|$?

- For 1. use stochastic coordinate ascent to solve $\|A\|^2 = \max_{\|v\|=1} \|Av\|^2$ with adjoint sampling:

$$v^{k+1/2} = v^k + \tau_k \langle Av^k, Ax\rangle x, \quad v^{k+1} = \frac{v^{k+1/2}}{\|v^{k+1/2}\|}$$

- Linesearch more difficult, but possible...
- For 2. use stochastic gradient ascent to solve $\|A - V\| = \max_{\|u\|=\|v\|=1} \langle u, (A-V)v\rangle = \max_{\|u\|=\|v\|=1} \left[\langle u, Av\rangle - \langle V^T u, v\rangle\right]$ by

$$v^{k+1/2} = v^k + \tau_k \left(\langle u^k, Ax\rangle x - V^T u^k\right), \qquad u^{k+1/2} = u^k + \tau_k \left(Av^k - \langle V^T y, v^k\rangle y\right)$$

$$v^{k+1} = \frac{v^{k+1/2}}{\|v^{k+1/2}\|}, \qquad u^{k+1} = \frac{u^{k+1/2}}{\|u^{k+1/2}\|},$$

Random descent

Lemma

The minimum of $\tau \mapsto \frac{1}{2}\|A(v^k + \tau x) - b\|^2$ *is attained at*

$$\tau_k = \begin{cases} -\frac{\langle Av^k - b, Ax\rangle}{\|Ax\|^2} & Ax \neq 0, \\ 0 & Ax = 0. \end{cases}$$

Does need neither $\|A\|$ nor A^T!
Gives *random descent* method (RD)

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

- Similar ideas: Random pursuit [Stich, Muller, Gartner, 2013], [Nesterov, Spokoiny, 2017]

Theorem

Let ν^k be generated by RD. Then it holds

$$\mathbf{E}(\|A\nu^{k-1} - b\|^2) = \|A\nu^k - b\|^2 - \langle A^T(A\nu^k - b), \mathbf{E}\left(\tfrac{xx^T}{\|Ax\|^2}\right)(A^T(A\nu^k - b)\rangle.$$

Convergence hinges on the spectral properties of

$$M := \mathbf{E}\left(\tfrac{xx^T}{\|Ax\|^2}\right) \in \mathbf{R}^{d\times d}$$

(if exists!). Simple and bad estimate:

$$\frac{xx^T}{\|Ax\|^2} \geq \frac{1}{\|A\|^2}\frac{xx^T}{\|x\|^2} \rightsquigarrow M \succcurlyeq \frac{1}{d\|A\|^2} I_d.$$

Gives only

$$\mathbf{E}(\|A\nu^{k+1} - b\|^2) \leq \left(1 - \tfrac{\sigma_{\min}(A)^2}{d\|A\|^2}\right)\|A\nu^k - b\|^2$$

Better results for specific distributions: Random coordinate vectors

- Gives randomized coordinate descent [Luenberger 1984, Leventhal, Lewis 2010]
- $\mathbf{P}(x = \sqrt{d}e_k) = \frac{1}{d}$, $a_k = Ae_k$

$$M = \sum_{k=1}^{d} \frac{1}{d} \frac{e_k e_k^T}{\|a_k\|^2} = \frac{1}{d}\operatorname{diag}(\|a_1\|^{-2}, \ldots, \|a_d\|^{-2}) \succcurlyeq \frac{1}{d \max_k \|a_k\|^2} I_d,$$

- Gives convergence

$$\mathbf{E}(\|Av^{k+1} - b\|^2) \leq \left(1 - \frac{\sigma_{\min}(A)^2}{d \max_k \|a_k\|^2}\right) \|Av^k - b\|^2.$$

$\max_k \|a_k\| \leq \max_{\|x\|=1} \|Ax\| = \|A\| \rightsquigarrow$ better than general bound

- Improved rate by choosing $\mathbf{P}(x = \sqrt{d}e_k) = \frac{\|a_k\|^2}{\|A\|_F^2}$ (precompute $\|A\|_F$!) leads to

$$\mathbf{E}(\|Av^{k+1} - b\|^2) \leq \left(1 - \frac{\sigma_{\min}(A)^2}{\|A\|_F^2}\right) \|Av^k - b\|^2 \qquad \text{[Leventhal, Lewis 2010]}$$

Better results for specific distributions: Standard normal vectors

- Here A^TA and M have same orthonormal eigenbasis (u_i) and the eigenvalues of M are

$$\mu_i = \lambda_i(M) = \mathbf{E}\left(\frac{\langle u_i, x\rangle}{\sum_{j=1}^d \lambda_j \langle u_j, x\rangle}\right) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbf{R}^d} \frac{z_i^2}{\lambda_1 z_1^2 + \cdots + \lambda_d z_d^2} e^{-\|z\|^2/2} \mathrm{d}z$$

$$\geq \frac{1}{2\|A\|_F^2} \frac{\Gamma(d/2)}{\Gamma((d+1)/2)} \approx \frac{1}{2\sqrt{2d}\|A\|_F^2}$$

- Gives convergence

$$\mathbf{E}(\|Av^{k+1} - b\|^2) \leq \left(1 - \frac{\sigma_{\min}(A)^2}{2\sqrt{2d}\|A\|_F^2}\right) \|Av^k - b\|^2$$

Experiments for consistent systems

- Comparison of SGDAS, RD (with Rademacher vectors), 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 $\hat{\nu}$ with normally distributed entries.

	$\frac{\|A\nu-b\|}{\|b\|}$	$\|\nu\|$	time (s)
SGDAS	9.68e-01	5.27e-01	8.67e-01
RD	**9.99e-03**	3.30e+01	3.76e-01
TFQMR	4.76e-02	3.35e+01	4.76e-01
CGS	**9.99e-03**	3.30e+01	**3.27e-04**

(a) Size of A: 300×1200, density of A: 0.1.

	$\frac{\|A\nu-b\|}{\|b\|}$	$\|\nu\|$	time (s)
SGDAS	8.72e-01	1.91e+00	6.89e-01
RD	**9.95e-03**	1.70e+01	**2.77e-01**
TFQMR	2.15e+00	2.37e+02	4.67e-01
CGS	4.23e+01	2.97e+03	7.52e-01

(b) Size of A: 1200×300, density of A: 0.1.

Experiments for consistent systems

- Comparison of SGDAS, RD (with Rademacher vectors), TFQMR and CGS.
- Stopped at relative tolerance of 10^{-5} or after 500000 iterations.
- Random sparse matrices A with normally distributed entries and random solution vectors $\hat{v}$ with normally distributed entries.

	$\frac{\|Av-b\|}{\|b\|}$	$\|v\|$	time (s)
SGDAS	1.16e-02	1.07e+01	8.56e+00
RD	**1.00e-05**	1.10e+01	**1.78e+00**
TFQMR	1.03e+00	1.23e+01	1.07e+01
CGS	4.33e+17	1.51e+22	1.34e+01

(c) Size of A: 200×100, density of A: 0.02.

	$\frac{\|Av-b\|}{\|b\|}$	$\|v\|$	time (s)
SGDAS	9.45e-03	1.05e+01	9.38e+00
RD	9.99e-06	1.07e+01	3.88e-01
TFQMR	**5.41e-07**	6.60e+01	3.33e-02
CGS	4.67e-06	6.60e+01	**2.12e-02**

(d) Size of A: 150×100, density of A: 0.1.

Convergence along singular vectors

- Consider $b = A\hat{\nu} + \eta$ and let $\{u_i\}$ be right singular vectors of A for singular values σ_i
- Simple observation for the Landweber iteration with stepsize ω:

$$\langle \nu^{k+1} - \hat{\nu}, u_i\rangle = (1 - \omega\sigma_i^2)^k \langle \nu^0 - \hat{\nu}, u_i\rangle + \tfrac{1-(1-\omega\sigma_i^2)^k}{\sigma_i}\langle \eta, u_i\rangle$$

$\rightsquigarrow$ faster decay of σ_i is larger (Similar for Kaczmarz [Jia, Jin, Lu 2017], [Steinerberger 2021])

- For random descent with standard normal directions:

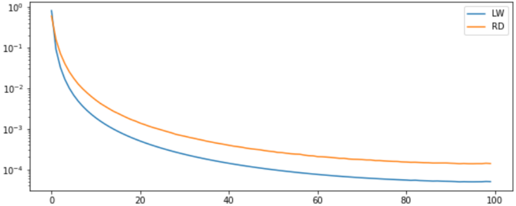
$$\mathbf{E}(\langle \nu^{k+1} - \hat{\nu}, u_i\rangle) = (1 - \mu_i\sigma_i^2)^k \langle \nu^0 - \hat{\nu}, u_i\rangle + \tfrac{1-(1-\mu_i\sigma_i^2)^k}{\sigma_i}\langle \eta, u_i\rangle$$

$\rightsquigarrow$ even faster decay if $\mu_i > 1/\omega$

- Random descent has advantage if $\nu^0 - \hat{\nu}$ is rough and the μ_i's are large

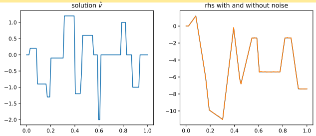
“Inverse integration”

- $A = \begin{bmatrix} 1 & & \\ \vdots & \ddots & \\ 1 & \cdots & 1 \end{bmatrix}$, $d = 100$
- $\|A\|^{-2}\sigma_i^2$ (for Landweber) vs. $\mu_i\sigma_i^2$ (for random descent)

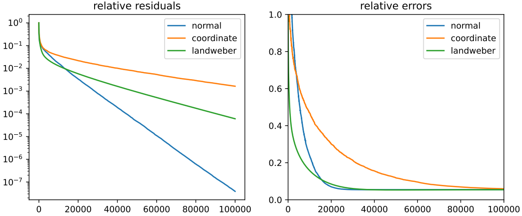


- Errors in higher singular modes decay way faster for random descent than for Landeweber

“Inverse integration” with rough data



- Convergence of error and residual:



- Stop by Morozov: RD after 36.256 iterations with 5.4% error, Landweber after 22,045 iterations with 7.7% error

Extension to non-linear least squares

- Consider $F : \mathbf{R}^d \to \mathbf{R}^m$ and $\min_{\mathbf{R}^d} \frac{1}{2}\|F(v) - b\|^2$.
- Landweber methods is

$$v^{k+1} = v^k - \tau DF(v^k)^T(F(v^k) - b)$$

  Needs transpose of derivative
- Adjoint sampling/stochastic Landweber + finite difference approximation

$$v^{k+1} = v^k - \tau\langle F(v^k) - b, DF(v^k)x\rangle x \approx v^k - \tau\langle F(v^k) - b, F(v^k + x) - F(v^k)\rangle x$$

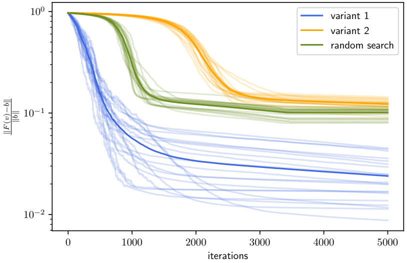
  Only need forward evaluations of F!
- Related: Random search [Polyak, 1987]. Minimize Φ by

$$v^{k+1} = v^k - \frac{\gamma_k}{\alpha_k}\left[\Phi(v^k + \alpha_k u) - \Phi(v^k)\right]u, \qquad u \sim \mathrm{Unif}(S^{d-1})$$

  Converges for γ_k small enough and $\alpha_k \to 0$.

A non-linear Hammerstein equation

- Consider $F(v)(s) = \int_0^1 |s-t| v(t)^3 \mathrm{d}t$, discretized to $F : \mathbf{R}^d \to \mathbf{R}^d$, $d = 200$, $\tau = 0.5/d$.
- Compare three variants:
  - Variant 1: $v^{k+1} = v^k - \tau\langle F(v^k) - b, F(v^k + x) - F(v^k)\rangle x$
  - Variant 2: $v^{k+1} = v^k - \langle F(v^k) - b, F(v^k + \tau x) - F(v^k)\rangle x$
  - Random search for $\Phi(v) = \frac{1}{2}\|F(v) - b\|^2$, $\gamma_k \equiv \gamma = 2$, $\alpha_k = 0.99^k$



Conclusion

- Transpose-free solution of least squares problems is possible by random descent and adjoint sampling
- Random descent even competitive with other transpose free methods like TFQMR and CGS
- Choice of distributions of directions influences convergence speed
- Coordinate descent as special case
- Random descent has some advantage for ill-posed problems with rough solutions
- Possible extensions:
  - Proximal methods with adjoint sampling
  - Isotropic sampling works, but problem adapted distributions may by better