Preconditioner Design via the Bregman Divergence

Joint work with Martin S. Andersen

Computational Mathematics for Data Science

Andreas Bock
17^{th} of November 2023

Technical University of Denmark

Problem setup

Find a solution to the following $n \times n$ linear system:

$$Sx = (A + B)x = b \tag{1}$$

- $A = QQ^*$ Hermitian positive definite, $x \mapsto Q^{-1}x$ known
- B Hermitian positive semidefinite

Problem setup

Find a solution to the following $n \times n$ linear system:

$$Sx = (A + B)x = b \tag{1}$$

- $A = QQ^*$ Hermitian positive definite, $x \mapsto Q^{-1}x$ known
- B Hermitian positive semidefinite

Motivating example: variational data assimilation

$$S = \underbrace{\mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L}}_{A} + \underbrace{\mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H}}_{B},$$

A is a forward model term ($\mathbf{L}$ is the model), B stems from an observation operator $\mathbf{H}$. $\mathbf{D}$ and $\mathbf{R}$ are (ill-conditioned) covariance matrices.

Problem setup

Find a solution to the following $n \times n$ linear system:

$$Sx = (A + B)x = b \tag{1}$$

- $A = QQ^*$ Hermitian positive definite, $x \mapsto Q^{-1}x$ known
- B Hermitian positive semidefinite

Motivating example: variational data assimilation

$$S = \underbrace{\mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L}}_{A} + \underbrace{\mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H}}_{B},$$

A is a forward model term ($\mathbf{L}$ is the model), B stems from an observation operator $\mathbf{H}$. $\mathbf{D}$ and $\mathbf{R}$ are (ill-conditioned) covariance matrices.

Question: what is the best preconditioner for (1) of the form

$$P = A + X, \quad \operatorname{rank}(X) \le r < n \quad ?$$

Preconditioned iterative methods

Situation

- S cannot be factorised directly but $x \mapsto Sx$ is available.
- Solutions to $Sx = b$ are sought via iterative methods e.g. the preconditioned conjugate gradient (PCG) method.

Preconditioned iterative methods

Situation

- S cannot be factorised directly but $x \mapsto Sx$ is available.
- Solutions to $Sx = b$ are sought via iterative methods e.g. the preconditioned conjugate gradient (PCG) method.

Preconditioned iterative methods

- Transform $Sx = b$ into:
$$P^{-1}Sx = P^{-1}b.$$
- Construction and application of P^{-1} must be cheap.
- Works well if $P \approx S$, generally we seek $\kappa(P^{-1}S) < \kappa(S)$.

Preconditioned iterative methods

Situation

- S cannot be factorised directly but $x \mapsto Sx$ is available.
- Solutions to $Sx = b$ are sought via iterative methods e.g. the preconditioned conjugate gradient (PCG) method.

Preconditioned iterative methods

- Transform $Sx = b$ into:
$$P^{-1}Sx = P^{-1}b.$$
- Construction and application of P^{-1} must be cheap.
- Works well if $P \approx S$, generally we seek $\kappa(P^{-1}S) < \kappa(S)$.

...but what does "$\approx$" mean?

Obvious discrepancy measures include $\|P - S\|_2$, $\|P - S\|_F$, ...

Bregman log determinant matrix divergence

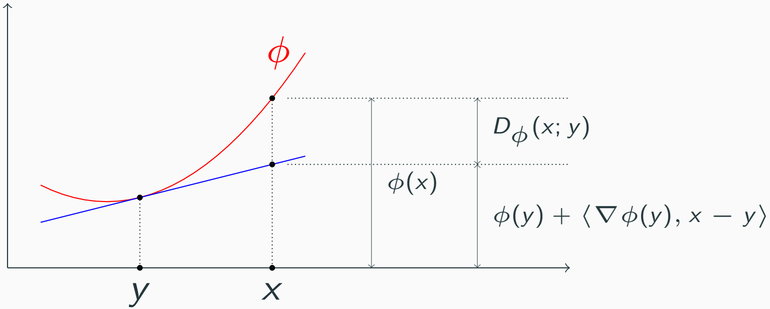
A proper and strictly convex function $\phi \in \mathrm{C}^1$ defines a **Bregman matrix divergence** $D_\phi : \operatorname{dom}\phi \times \operatorname{ri}\operatorname{dom}\phi \to [0, \infty)$:

$$D_\phi(X, Y) = \phi(X) - \phi(Y) - \langle \nabla\phi(Y), (X - Y) \rangle.$$

Bregman log determinant matrix divergence

A proper and strictly convex function $\phi \in C^1$ defines a **Bregman matrix divergence** $D_\phi : \operatorname{dom}\phi \times \operatorname{ri}\operatorname{dom}\phi \to [0, \infty)$:

$$D_\phi(X, Y) = \phi(X) - \phi(Y) - \langle \nabla\phi(Y), (X - Y)\rangle.$$



Bregman log determinant matrix divergence

A proper and strictly convex function $\phi \in \mathrm{C}^1$ defines a **Bregman matrix divergence** $D_\phi : \operatorname{dom}\phi \times \operatorname{ri}\operatorname{dom}\phi \to [0, \infty)$:

$$D_\phi(X, Y) = \phi(X) - \phi(Y) - \langle \nabla\phi(Y), (X - Y) \rangle.$$

Bregman log determinant matrix divergence

A proper and strictly convex function $\phi \in \mathrm{C}^1$ defines a **Bregman matrix divergence** $D_\phi : \operatorname{dom}\phi \times \operatorname{ri}\operatorname{dom}\phi \to [0, \infty)$:

$$D_\phi(X, Y) = \phi(X) - \phi(Y) - \langle \nabla\phi(Y), (X - Y)\rangle.$$

$$\begin{array}{lcl} \phi(X) = \frac{1}{2}\|X\|_F^2 & \to & D_F(X, Y) = \frac{1}{2}\|X - Y\|_F^2 \\ \phi(X) = -\log\det(X) & \to & D_B(X, Y) = \operatorname{trace}(XY^{-1}) - \log\det(XY^{-1}) - n \end{array}$$

Bregman log determinant matrix divergence

A proper and strictly convex function $\phi \in C^1$ defines a **Bregman matrix divergence** $D_\phi : \operatorname{dom}\phi \times \operatorname{ri}\operatorname{dom}\phi \to [0, \infty)$:

$$D_\phi(X, Y) = \phi(X) - \phi(Y) - \langle \nabla\phi(Y), (X - Y)\rangle.$$

$$\begin{array}{lcl} \phi(X) = \frac{1}{2}\|X\|_F^2 & \to & D_F(X, Y) = \frac{1}{2}\|X - Y\|_F^2 \\ \phi(X) = -\log\det(X) & \to & D_B(X, Y) = \operatorname{trace}(XY^{-1}) - \log\det(XY^{-1}) - n \end{array}$$

Properties

- $D_\phi(X, Y) = 0 \Leftrightarrow X = Y$,
- *Nonnegativity*: $D_\phi(X, Y) \geq 0$,
- *Convexity*: $X \to D_\phi(X, Y)$ is convex.
- In addition, D_B is invariant under congruence transformations:

For invertible $\mathbf{M}$ we have $D_B(X, Y) = D_B(\mathbf{M}^* X \mathbf{M}, \mathbf{M}^* Y \mathbf{M})$.

Preconditioners as Bregman projections

Recall $S = A + B$**,** $A = QQ^*$

Preconditioners as Bregman projections

Recall $S = A + B$, $A = QQ^*$

Candidates: $P = A + X = Q(I + Q^{-1}XQ^{-*})Q^*$, **where** $\operatorname{rank}(X) \leq r < n$

Preconditioners as Bregman projections

Recall $S = A + B$**,** $A = QQ^*$

Candidates: $P = A + X = Q(I + Q^{-1}XQ^{-*})Q^*$**, where** $\operatorname{rank}(X) \leq r < n$

We solve:

$$\begin{aligned} \underset{W \in \mathbb{H}^n_+}{\text{minimise}} \quad & D_B(P, S) = \operatorname{trace}(PS^{-1}) - \log\det(PS^{-1}) - n \\ \text{s.t.} \quad & P = Q(I + W)Q^* \quad \text{(change of var. from } X \text{ to } W\text{)} \\ & \operatorname{rank}(W) \leq r \end{aligned}$$

Preconditioners as Bregman projections

Recall $S = A + B$, $A = QQ^*$

Candidates: $P = A + X = Q(I + Q^{-1}XQ^{-*})Q^*$, **where** $\operatorname{rank}(X) \leq r < n$

We solve:

$$\begin{aligned} \underset{W \in \mathbb{H}_+^n}{\text{minimise}} \quad & D_B(P, S) = \operatorname{trace}(PS^{-1}) - \log\det(PS^{-1}) - n \\ \text{s.t.} \quad & P = Q(I + W)Q^* \quad (\text{change of var. from } X \text{ to } W) \\ & \operatorname{rank}(W) \leq r \end{aligned}$$

Invariance to the rescue:

$$\begin{aligned} D_B(P, S) &= D_B(Q(I + W)Q^*, Q(I + Q^{-1}BQ^{-*})Q^*) \\ &= D_B(I + W, I + Q^{-1}BQ^{-*}) \end{aligned}$$

Preconditioners as Bregman projections

Recall $S = A + B$**,** $A = QQ^*$

Candidates: $P = A + X = Q(I + Q^{-1}XQ^{-*})Q^*$**, where** $\operatorname{rank}(X) \leq r < n$

We solve:

$$\begin{aligned} \underset{W \in \mathbb{H}_+^n}{\text{minimise}} \quad & D_B(P, S) = \operatorname{trace}(PS^{-1}) - \log\det(PS^{-1}) - n \\ \text{s.t.} \quad & P = Q(I + W)Q^* \quad (\text{change of var. from } X \text{ to } W) \\ & \operatorname{rank}(W) \leq r \end{aligned}$$

Invariance to the rescue:

$$\begin{aligned} D_B(P, S) &= D_B(Q(I + W)Q^*, Q(I + Q^{-1}BQ^{-*})Q^*) \\ &= D_B(I + W, I + Q^{-1}BQ^{-*}) \end{aligned}$$

Reduced problem:

$$\begin{aligned} \underset{W \in \mathbb{H}_+^n}{\text{minimise}} \quad & D_B(I + W, I + Q^{-1}BQ^{-*}) \\ \text{s.t.} \quad & \operatorname{rank}(W) \leq r. \end{aligned}$$

Summary of theoretical results

Summary of theoretical results

Theorem
Let G_r be a rank r truncated SVD of $G = Q^{-1}BQ^{-*}$.

$$P^\star = Q(I + G_r)Q^*$$

is a minimiser of $D_B(P, S)$ over the set of preconditioners of the form $P = A + X$, $\operatorname{rank}(X) \leq r$.

Summary of theoretical results

Theorem

Let G_r be a rank r truncated SVD of $G = Q^{-1}BQ^{-*}$.

$$P^\star = Q(I + G_r)Q^*$$

is a minimiser of $D_B(P, S)$ over the set of preconditioners of the form $P = A + X$, $\text{rank}(X) \leq r$.

Note, in general, $P^\star \neq A + B_r$ (''$=$'' holds when, e.g., $A = \sigma^2 I$)

Summary of theoretical results

Theorem
Let G_r be a rank r truncated SVD of $G = Q^{-1}BQ^{-*}$.

$$P^\star = Q(I + G_r)Q^*$$

is a minimiser of $D_B(P, S)$ over the set of preconditioners of the form $P = A + X$, $\operatorname{rank}(X) \leq r$.

Note, in general, $P^\star \neq A + B_r$ ("$=$" holds when, e.g., $A = \sigma^2 I$)

Theorem
When $\operatorname{rank}(B) < n$, G_r is a minimiser of the problem

$$\begin{aligned} \underset{X \in \mathbb{H}_+^n}{\text{minimise}} \quad & \kappa_2(P^{-\frac{1}{2}} S P^{-\frac{1}{2}}) \\ \text{s.t.} \quad & P = Q(I + X)Q^* \\ & \operatorname{rank}(X) \leq r. \end{aligned}$$

Digression: which parts of $G = Q^{-1}BQ^{-*}$ do we amputate?

Digression: which parts of $G = Q^{-1}BQ^{-*}$ do we amputate?

Hermitian rank r approximations W such that $\|G - W\| < \epsilon$

Digression: which parts of $G = Q^{-1}BQ^{-*}$ do we amputate?

Hermitian rank r approximations W such that $\|G - W\| < \epsilon$

- **Truncated SVD:** $$G_r = U_r \Sigma_r U_r^* \quad G = U \Sigma U^*$$
- **Randomised SVD:**
$$G_{\mathrm{RSVD}} = \Theta\Theta^\top G \Theta\Theta^\top$$
where $\Theta R = \Omega \in \mathbb{R}^{n\times r}$ (columns of Ω are Gaussian)
- **Nyström:** $$G_{\mathrm{Nys}} = G\Omega(\Omega^* G\Omega)^\dagger (G\Omega)^*$$

Digression: which parts of $G = Q^{-1}BQ^{-*}$ do we amputate?

Hermitian rank r approximations W such that $\|G - W\| < \epsilon$

- **Truncated SVD:** $G_r = U_r \Sigma_r U_r^*$ $\quad G = U\Sigma U^*$
- **Randomised SVD:**

$$G_{\mathrm{RSVD}} = \Theta\Theta^\top G \Theta\Theta^\top$$

  where $\Theta R = \Omega \in \mathbb{R}^{n\times r}$ (columns of Ω are Gaussian)

- **Nyström:** $G_{\mathrm{Nys}} = G\Omega(\Omega^* G\Omega)^\dagger (G\Omega)^*$

Folklore: G_r is "better" than G_{Nys}, which is "better" than G_{RSVD}...

Digression: which parts of $G = Q^{-1}BQ^{-*}$ do we amputate?

Hermitian rank r approximations W such that $\|G - W\| < \epsilon$

- **Truncated SVD:** $$G_r = U_r \Sigma_r U_r^* \quad G = U\Sigma U^*$$
- **Randomised SVD:** $$G_{\mathrm{RSVD}} = \Theta\Theta^\top G \Theta\Theta^\top$$

  where $\Theta R = \Omega \in \mathbb{R}^{n\times r}$ (columns of Ω are Gaussian)
- **Nyström:** $$G_{\mathrm{Nys}} = G\Omega(\Omega^* G\Omega)^\dagger (G\Omega)^*$$

Folklore: G_r is "better" than G_{Nys}, which is "better" than G_{RSVD}...

Theorem
G_{Nys} is a minimiser of a *range-restricted* Bregman divergence:

$$\min_{W \in \mathbb{H}_+^n} \quad D(\Omega^* W \Omega, \Omega^* G \Omega)$$
$$\text{s.t.} \quad \operatorname{range} W \subseteq \operatorname{range} G\Omega.$$

Why does the Bregman divergence appear so useful?

Why does the Bregman divergence appear so useful?

By a Taylor expansion we have

$$D(X, X + \delta X) \approx \frac{1}{2}\mathrm{trace}(\delta X\, X^{-1}\delta X\, X^{-1}) = \frac{1}{2}g_X(\delta X, \delta X),$$

and $\mathcal{M} = (\mathbb{H}^n_{++}, g)$ is a Riemannian manifold.

Geometric insights

Why does the Bregman divergence appear so useful?

By a Taylor expansion we have

$$D(X, X+\delta X) \approx \frac{1}{2}\mathrm{trace}(\delta X\, X^{-1}\delta X\, X^{-1}) = \frac{1}{2} g_X(\delta X, \delta X),$$

and $\mathcal{M} = (\mathbb{H}^n_{++}, g)$ is a Riemannian manifold.

Theorem
$P^\star = Q(I + G_r)Q^*$ minimises the Riemannian distance to S given by

$$d_{\mathcal{M}}(P, S) = \| \operatorname{Log}\left(P^{-\frac{1}{2}} S P^{-\frac{1}{2}}\right)\|_2^2$$

among matrices of the form $Q(I + X)Q^*$ for some $X \in \mathbb{H}^n_+$ with $\mathrm{rank}(X) \leq r$.

Why does the Bregman divergence appear so useful?

By a Taylor expansion we have

$$D(X, X + \delta X) \approx \frac{1}{2}\mathrm{trace}(\delta X\, X^{-1} \delta X\, X^{-1}) = \frac{1}{2} g_X(\delta X, \delta X),$$

and $\mathcal{M} = (\mathbb{H}^n_{++}, g)$ is a Riemannian manifold.

Theorem
$P^\star = Q(I + G_r)Q^*$ minimises the Riemannian distance to S given by

$$d_{\mathcal{M}}(P, S) = \| \operatorname{Log}\left(P^{-\frac{1}{2}} S P^{-\frac{1}{2}}\right) \|_2^2$$

among matrices of the form $Q(I + X)Q^*$ for some $X \in \mathbb{H}^n_+$ with $\mathrm{rank}(X) \leq r$.

Many things to explore

Low-rank geodesic shooting algorithms, alternating projection algorithms, dually flat Riemannian structure, Stiefel manifold optimisation...

Application to variational data assimilation

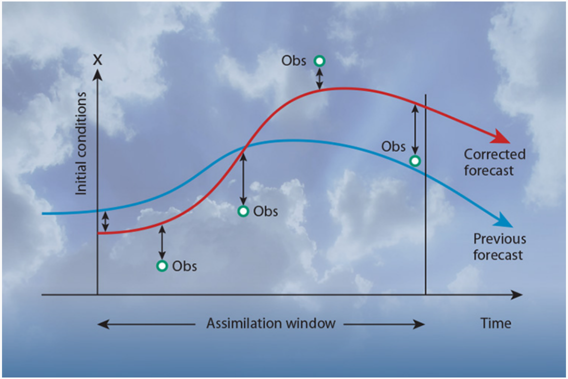


Image retrieved from the European Centre for Medium-Range Weather Forecasts (`www.ecmwf.int`)

Application to variational data assimilation

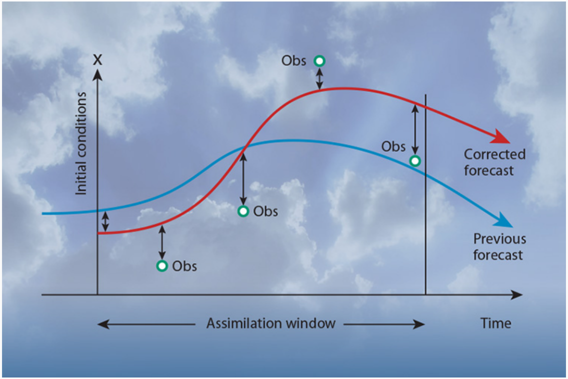


Image retrieved from the European Centre for Medium-Range Weather Forecasts (`www.ecmwf.int`)

$$J(x_0) = \underbrace{\frac{1}{2}(x_0 - x_0^B)^\top B^{-1}(x_0 - x_0^B)}_{\text{initial cond.}} + \underbrace{\frac{1}{2}\sum_{i=1}^{N}(x_i - \mathcal{M}_i(x_{i-1}))^\top Q_i^{-1}(x_i - \mathcal{M}_i(x_{i-1}))}_{\text{forward model}}$$

$$+ \underbrace{\frac{1}{2}\sum_{i=0}^{N}(y_i - \mathcal{H}_i(x_i))^\top R_i^{-1}(y_i - \mathcal{H}_i(x_i))}_{\text{match observations}}$$

Application to variational data assimilation

Gauss-Newton for weak constraint 4D VAR

Application to variational data assimilation

Gauss-Newton for weak constraint 4D VAR

At each GN step, we solve for the increment $\delta\mathbf{x}$ by inverting the Hessian of J_{GN}:

$$S = \underbrace{\mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L}}_{A} + \underbrace{\mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H}}_{B},$$

$$\mathbf{D} = \begin{bmatrix} B & & & \\ & Q_1 & & \\ & & \ddots & \\ & & & Q_N \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} I & & \\ -M_1 & I & \\ & \ddots & \ddots \\ & -M_n & I \end{bmatrix}, \quad \begin{aligned} \mathbf{R} &= \mathbf{blkdiag}\,(R_0, \ldots, R_N), \\ \mathbf{H} &= \mathbf{blkdiag}\,(H_0, \ldots, H_N). \end{aligned}$$

Gauss-Newton for weak constraint 4D VAR

At each GN step, we solve for the increment $\delta\mathbf{x}$ by inverting the Hessian of J_{GN}:

$$S = \underbrace{\mathbf{L}^\top \mathbf{D}^{-1}\mathbf{L}}_{A} + \underbrace{\mathbf{H}^\top \mathbf{R}^{-1}\mathbf{H}}_{B},$$

$$\mathbf{D} = \begin{bmatrix} B & & & \\ & Q_1 & & \\ & & \ddots & \\ & & & Q_N \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} I & & & \\ -M_1 & I & & \\ & & \ddots & \\ & & -M_n & I \end{bmatrix}, \quad \begin{array}{l} \mathbf{R} = \mathbf{blkdiag}\,(R_0, \ldots, R_N), \\ \mathbf{H} = \mathbf{blkdiag}\,(H_0, \ldots, H_N). \end{array}$$

Example: assimilating the heat equation $\partial_t u = \Delta u$

$n = 10^5$, $s = 1000$ (spatial resolution), $N = 100$ (time steps), $\Delta t = 10^{-4}$ (step size)

$\mathrm{rank}(B) = n/2$ (we only observe half of the state at each time step)

$r \in \{500, 2000, 4000\}$ (about 0.05%, 2% and 4% of n, respectively)

Gauss-Newton for weak constraint 4D VAR

At each GN step, we solve for the increment $\delta\mathbf{x}$ by inverting the Hessian of J_{GN}:

$$S = \underbrace{\mathbf{L}^\top \mathbf{D}^{-1} \mathbf{L}}_{A} + \underbrace{\mathbf{H}^\top \mathbf{R}^{-1} \mathbf{H}}_{B},$$

$$\mathbf{D} = \begin{bmatrix} B & & & \\ & Q_1 & & \\ & & \ddots & \\ & & & Q_N \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} I & & & \\ -M_1 & I & & \\ & & \ddots & \\ & & -M_n & I \end{bmatrix}, \quad \begin{array}{c} \mathbf{R} = \mathbf{blkdiag}\,(R_0, \ldots, R_N), \\ \mathbf{H} = \mathbf{blkdiag}\,(H_0, \ldots, H_N). \end{array}$$

Example: assimilating the heat equation $\partial_t u = \Delta u$

$n = 10^5$, $s = 1000$ (spatial resolution), $N = 100$ (time steps), $\Delta t = 10^{-4}$ (step size)

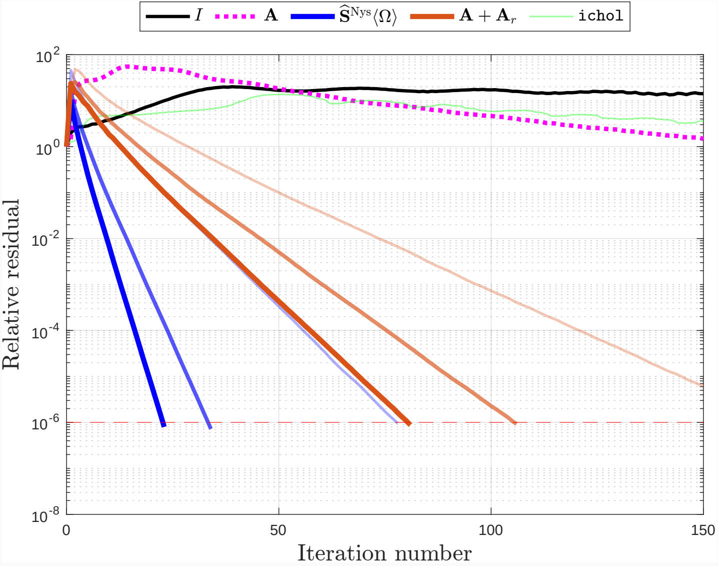
$\mathrm{rank}(B) = n/2$ (we only observe half of the state at each time step)

$r \in \{500, 2000, 4000\}$ (about 0.05%, 2% and 4% of n, respectively)

We compare the following preconditioners

$$P = A, \quad P = A + B_r, \quad \text{and} \quad P = Q(I + G_r)Q^\top.$$

Application to variational data assimilation



$r \in \{500, 2000, 4000\}$

Summary

Insights

- The Bregman divergence appears useful for studying preconditioners.

Summary

Insights

- The Bregman divergence appears useful for studying preconditioners.
- Importance of **invariance** cannot be understated.

Summary

Insights

- The Bregman divergence appears useful for studying preconditioners.
- Importance of **invariance** cannot be understated.
- Nyström can be derived using the Bregman divergence, where to next?
- Try it: `pip install scaled-preconditioners`

Summary

Insights

- The Bregman divergence appears useful for studying preconditioners.
- Importance of **invariance** cannot be understated.
- Nyström can be derived using the Bregman divergence, where to next?
- Try it: `pip install scaled-preconditioners`

Generalisations and future work

Summary

Insights

- The Bregman divergence appears useful for studying preconditioners.
- Importance of **invariance** cannot be understated.
- Nyström can be derived using the Bregman divergence, where to next?
- Try it: `pip install scaled-preconditioners`

Generalisations and future work

- What if you don't know the $A + B$ structure?

Summary

Insights

- The Bregman divergence appears useful for studying preconditioners.
- Importance of **invariance** cannot be understated.
- Nyström can be derived using the Bregman divergence, where to next?
- Try it: `pip install scaled-preconditioners`

Generalisations and future work

- What if you don't know the $A + B$ structure?
- Allowing indefiniteness of B: *coming soon to an arXiv near you!*

Summary

Insights

- The Bregman divergence appears useful for studying preconditioners.
- Importance of **invariance** cannot be understated.
- Nyström can be derived using the Bregman divergence, where to next?
- Try it: `pip install scaled-preconditioners`

Generalisations and future work

- What if you don't know the $A + B$ structure?
- Allowing indefiniteness of B: *coming soon to an arXiv near you!*
- Bounded (or other) divergences (numerical stability, more geometric insights)...
- **Big picture:** studying the *geometry* of preconditioners.

References

Dhillon, Inderjit S and Joel A Tropp (2008). **"Matrix nearness problems with Bregman divergences".** In: *SIAM Journal on Matrix Analysis and Applications* 29(4), pp. 1120–1146.

Kulis, Brian, Mátyás A Sustik, and Inderjit S Dhillon (2009). **"Low-Rank kernel learning with Bregman matrix divergences.".** In: *Journal of Machine Learning Research* 10(2).

Amari, Shun-ichi (2016). **Information geometry and its applications.** Vol. 194. Springer.

Martinsson, Per-Gunnar and Joel A Tropp (2020). **"Randomized numerical linear algebra: Foundations and algorithms".** In: *Acta Numerica* 29, pp. 403–572.

Tabeart, Jemima M. and John W. Pearson (2021). **"Saddle point preconditioners for weak-constraint 4D-Var".** In: *arXiv preprint arXiv:2105.06975*.

Bock, Andreas and Martin S Andersen (2023). **"Preconditioner Design via the Bregman Divergence".** In: *arXiv preprint arXiv:2304.12162*.

Thank you to everyone for coming to our workshop! ☺