Connecting regularization across scales for hybrid solutions of ill-posed problems

Rosemary Renaut\textsuperscript{1}  Anthony Helmstetter\textsuperscript{1}  Saeed Vatankhah\textsuperscript{2}

\textsuperscript{1}: School of Mathematical and Statistical Sciences, Arizona State University, renaut@asu.edu, anthony.helmstetter@asu.edu

\textsuperscript{2}: Institute of Geophysics, University of Tehran, svatan@ut.ac.ir

SIAM Houston Imaging Symposium
Motivation Example: Large Scale Gravity Inversion

Background: SVD for the small scale
- Standard Approaches to Estimate Regularization Problem
- UPRE is a good regularization parameter estimation tool

Methods for the Large Scale: Approximating the SVD
- Krylov: Golub Kahan Bidiagonalization - LSQR
- Parameter estimation on the projected problem
  - Details: UPRE is a good estimator for LSQR [RVA17]
- Randomized SVD: Convergence and Regularization

Parameter Estimation

Simulations
- One Dimension Contrasting the RSVD and LSQR: Some Trivial Experiments

Simulations: Two dimensional Examples
- Undersampled 3D gravity data: approximate $L_1$ regularization

Conclusions
Motivation Example: Large Scale 3D Gravity Inversion

Observation point $\mathbf{r} = (x, y, z)$

Vertical gravitational attraction $g(\mathbf{r})$

$$g(\mathbf{r}) = \Gamma \int d\Omega \varrho(\mathbf{r}') \frac{\mathbf{r}' - \mathbf{r}}{|\mathbf{r}' - \mathbf{r}|^3} d\Omega'$$

Density $\varrho(\mathbf{r}')$ at $\mathbf{r}' = (x', y', z')$

Newton gravitational constant: $\Gamma$

Aim: Given surface observations $g_{ij}$ find volume density $\varrho_{ijk}$

Underdetermined, measurements 5500, unknowns 66000

Practical Approaches to Solve Large Scale are needed
Consider general discrete problem

\[ Ax = b, \quad A \in \mathbb{R}^{m \times n}, \quad b \in \mathbb{R}^m, \quad x \in \mathbb{R}^n. \]

Singular value decomposition (SVD) of \( A \) rank \( r \)

\[
A = U \Sigma V^T = \sum_{i=1}^{r} u_i \sigma_i v_i^T, \quad \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r).
\]

gives expansion for the solution

\[
x = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i
\]

Truncation: \( k < r \)

\[
x = \sum_{i=1}^{k} \frac{u_i^T b}{\sigma_i} v_i
\]

Filtering \( \gamma_i \)

\[
x = \sum_{i=1}^{r} \gamma_i(\alpha) \frac{u_i^T b}{\sigma_i} v_i
\]

Filtered Truncated

\[
x = \sum_{i=1}^{k} \gamma_i(\alpha) \frac{u_i^T b}{\sigma_i} v_i
\]
Tikhonov Regularization: regularization parameter $\alpha$

Filter Functions

$$\gamma_i(\alpha) = \frac{\sigma_i^2}{\sigma_i^2 + \alpha^2}, \quad i = 1 \ldots r,$$

Solves Standard Form

$$x(\alpha) = \arg\min_x \{ \|b - Ax\|^2 + \alpha^2 \|x\|^2 \}$$

Generalized Tikhonov has operator $L$

$$x(\alpha) = \arg\min_x \{ \|b - Ax\|^2 + \alpha^2 \|Lx\|^2 \}$$

Solve with standard form if $L$ invertible.

Requires automatic estimation of $\alpha$

How to solve and how to find $\alpha^{opt}$ for large scale problems?
Introduce $\phi_i(\alpha) = \frac{\alpha^2}{\alpha^2 + \sigma_i^2} = 1 - \gamma_i(\alpha)$, $i = 1 : r$, $\gamma_i = 0$, $i > k$.

**Unbiased Predictive Risk** : Minimize functional noise level $\sigma^2$

$$U_k(\alpha) = \sum_{i=1}^{k} \phi_i^2(\alpha)(u_i^T b)^2 - 2\sigma^2 \sum_{i=1}^{k} \phi_i(\alpha)$$

**GCV** : Minimize rational function $m^* = \min\{m, n\}$

$$G(\alpha) = \left( \sum_{i=1}^{m^*} \phi_i^2(\alpha)(u_i^T b)^2 \right) \left( \sum_{i=1}^{m^*} \phi_i^2(\alpha) \right)^{-2}$$

**How does $\alpha^{opt} = \arg\min F(\alpha)$ depend on $k$?**
Examples: $F(\alpha)$ Increasing truncation $k$. Noise level $\sigma^2 = .0001$

UPRE function $U_k(\alpha)$ with increasing $k$: unique minimum

GCV function $G_k(\alpha)$ with increasing $k$: non unique minimum
Theorem on UPRE for the FTSVD regularization

Coefficients of the data: \( \hat{b}_i = u_i^T b \)

Assumption: There exists \( \ell \) such that \( E(\hat{b}_i^2) = \sigma^2 \) for all \( i > \ell \), i.e. coefficients \( \hat{b}_i \) noise contaminated \( i > \ell \).

Define

\[ \alpha^{opt} = \operatorname{argmin} U(\alpha) \quad \text{and} \quad \alpha_k = \operatorname{argmin} U_k(\alpha), \]

Theorem

Suppose that \( k = \ell + p, p > 0 \), then the sequence \( \{\alpha_k\} \) is on the average increasing with \( \lim_{k \to r} \alpha_k = \alpha^{opt} \). Furthermore \( \{U_k(\alpha_k)\} \) is increasing, with \( \lim_{k \to r} U_k(\alpha_k) = U(\alpha^{opt}) \).

Observations:

1. Find \( \alpha_k \) for TSVD with \( k \) terms.
2. Determine optimal \( k \) as \( \alpha_k \) converges to \( \alpha^{opt} \).
Traditional: The LSQR iteration: Given $k$ defines range space

**LSQR** Let $\beta_1 := \|b\|_2$, and $e_1^{(k+1)}$ first column of $I_{k+1}$
Generate, lower bidiagonal $B_k \in \mathcal{R}^{(k+1) \times k}$, column orthonormal $H_{k+1} \in \mathcal{R}^{m \times (k+1)}$, $G_k \in \mathcal{R}^{n \times k}$

$$AG_k = H_{k+1}B_k, \quad \beta_1 H_{k+1}e_1^{(k+1)} = b.$$ 

**Projected Problem** on projected space: (standard Tikhonov)

$$w_k(\zeta_k) = \arg\min_{w \in \mathcal{R}^k} \{\|B_kw - \beta_1e_1^{(k+1)}\|_2^2 + \zeta_k^2\|w\|_2^2\}.$$ 

**Projected Solution** depends on $\zeta_k^{opt}$

$$x_k(\zeta_k^{opt}) = G_kw_k(\zeta_k^{opt}).$$

Generally: $\zeta_k^{opt} \neq \alpha^{opt}$

**Regularization** is on $\|w\|^2 = \|x\|^2$. Complicates generalized Tikhonov, $\|Lw\|^2 \neq \|Lx\|^2$
Regularization of the LSQR solution: Questions

(i) Determine optimal $k$: the size of the Krylov subspace. The choice of the subspace impacts the regularizing properties of the iteration: For large $k$ noise due to numerical precision and data error enters the projected space.

(ii) Determine optimal $\zeta_k$: How do regularization parameter techniques translate to the projected problem?

(iii) Relation optimal $\zeta_k$ and optimal $\alpha$: Given $k$ how well does optimal $\zeta_k$ for projected space yield optimal $\alpha$ for full space, or when is this the case? How does UPRE do as compared to GCV.

(iv) WGCV Note [CNO08] GCV regularization requires weighting parameter $\omega$
Regularization of the LSQR solution by UPRE [RVA17]

\[ U^{\text{full}}(\alpha) \] functional for original problem (depends on \( A, b \))

\[ U^{\text{proj}}(\zeta_k) \] functional for projected problem (depends on \( B_k \) and \( \beta_1 e_1^{(k+1)} \))

Is \( \alpha^{\text{opt}} \) relevant to \( \zeta_k^{\text{opt}} \) for the projected problem?

Theorem ([RVA17]: with assumptions on the approximation of the spectral space)

1. \( \alpha^{\text{opt}} \) for \( U^{\text{full}}(\alpha) \) can be estimated for projected problem
2. Minimizer of \( U^{\text{proj}}(\zeta_k) \) is minimizer of \( U^{\text{full}}(\zeta_k) \)
3. \( \zeta_k^{\text{opt}} \) depends on \( k \), \( \alpha^{\text{opt}} \) depends on \( m^* =: \min(m, n) \)
4. If \( k^* \) approx numerical rank \( A \), and right singular space is well-approximated \( \zeta_k^{\text{opt}} \approx \alpha^{\text{opt}} \) for \( K_{k^*}(A^T A, A^T b) \)

But when is the singular space approximation good enough
Approximation of the spectral space: \( A' = H_{k+1} B_k G_k^T \neq A \)

**Lemma** (Relates the singular spaces of \( A' \) and \( B_k \))

Let \( B_k = \tilde{U} \tilde{\Sigma} \tilde{V}^T \), \( H_{k+1} B_k G_k^T = U' \Sigma' V'^T \). If \( \sigma_i \neq \sigma_j \ \forall i \), then

\[
(H_{k+1} \tilde{U}) \tilde{\Sigma} (\tilde{V}^T G_k^T) = H_{k+1} B_k G_k^T,
\]

and, up to sign change,

\[
(H_{k+1} \tilde{U})(:, 1 : k) = U'(:, 1 : k) \quad (G_k \tilde{V})(:, 1 : k) = V'(:, 1 : k)
\]

Define \( \gamma_k = \| A - A' \| = \| \delta A \| \) then:

**Theorem** ([DDLT91]: how far is \( A' \) from \( A \)? \( (\sigma_i \neq \sigma_j) \))

\( u_i \) and \( u'_i \) (\( v_i \) and \( v'_i \)) are left (right) unit singular vectors of \( A \) and \( A' \). For \( \| \delta A \| \leq \gamma_k \), if \( 2 \gamma_k < \min_{i \neq j} |\sigma_i - \sigma_j| \), then

\[
\max(\sin \Theta(u_i, u'_i), \sin \Theta(v_i, v'_i)) \leq \frac{\gamma_k}{\min_{i \neq j} |\sigma_i - \sigma_j| - \gamma_k} \leq 1.
\]
Convergence: depends on the gap $\gamma_k$

**Theorem ([Jia16]: For fast decay of singular values)**

For $\ell: \hat{b}_\ell > \sigma_\ell$. Decay rate $\sigma_i = \zeta \rho^{-i}$, $\rho > 2$. Then $A' = H_{k+1}B_kG_k^T$ is near best rank $k$ approximation to $A$ for $k = 1, 2, \ldots, \ell$.

**Decay Rate** Results depend on decay rates

**Convergence** of UPRE thus depends on finding $k$ for which an optimal spectral approximation is found

**Truncation** With a given subspace a truncated space of size $k' < k$ provides an optimal approximation. Requires optimal determination of $k'$? (Not addressed in this talk)
Details: Calculating Unbiased Predictive Risk using $w_k(\alpha)$[RVA17]

Residual: $\mathbf{R}^{\text{full}}(\mathbf{x}_k) = A\mathbf{x}_k - \mathbf{b}$.

Influence Matrix $A(\alpha) = A(A^TA + \alpha^2 I)^{-1}A^T$

UPRE : Full problem

$$\alpha^{\text{opt}} = \arg\min_{\alpha} \left\{ \| \mathbf{R}^{\text{full}}(\mathbf{x}_k(\alpha)) \|_2^2 + 2 \text{Tr}(A(\alpha)) - m \right\} = \arg\min_{\alpha} \{ U^{\text{full}}(\alpha) \}$$

Using the projected solution for parameter $\alpha$ and

$$\text{Tr}((AG_k(\alpha))) = \text{Tr}(B_k(\alpha))$$

$$U^{\text{full}}(\alpha) = \| (AG_k(\alpha) - I_m) \mathbf{b} \|_2^2 + 2 \text{Tr}((AG_k(\alpha)) - m$$

$$= \| \beta_1(B_k(\alpha) - I_{k+1})e_1^{k+1} \|_2^2 + 2 \text{Tr}(B_k(\alpha)) - m$$

$\alpha^{\text{opt}}$ for $U^{\text{full}}(\alpha)$ can be estimated for projected problem
Deriving UPRE for the projected problem

**Is** $\alpha^{\text{opt}}$ **relevant to** $\zeta_k^{\text{opt}}$ **for the projected problem?**

**Noise in the right hand side**

For $b = b^{\text{true}} + \eta$, $\eta \sim \mathcal{N}(0, I_m)$

$$\beta_1 e_1^{k+1} = H_{k+1}^T b = H_{k+1}^T b^{\text{true}} + H_{k+1}^T \eta.$$ 

**Noise in projected right hand side**

$H_{k+1}^T \eta \sim \mathcal{N}(0, I_{k+1})$

Immediately

$$U_{\text{proj}}(\zeta_k) = \|\beta_1 (B_k(\zeta_k) - I_{k+1}) e_1^{(k+1)}\|_2^2 + 2 \text{Tr}(B_k(\zeta_k)) - (k + 1)$$

$$= U_{\text{full}}(\zeta_k) + m - (k + 1).$$

**Minimizer of** $U_{\text{proj}}(\zeta_k)$ **is minimizer of** $U_{\text{full}}(\zeta_k)$
$\zeta_k^{\text{opt}}$ calculated for projected problem may not yield $\alpha^{\text{opt}}$ on full problem.

$\zeta_k^{\text{opt}}$ depends on $k$, $\alpha^{\text{opt}}$ depends on $m^* =: \min(m, n)$.

**Trace Relations** By linearity and cycling.

\[
\text{Tr}(A(\alpha)) = \text{Tr}(A(A^T A + \alpha^2 I_n)^{-1} A^T) = m^* - \alpha^2 \sum_{i=1}^{m^*} (\sigma_i^2 + \alpha^2)^{-1}
\]

\[
\text{Tr}(B_k(\zeta_k)) = k - \zeta_k^2 \sum_{i=1}^{k} (\gamma_i^2 + \zeta_k^2)^{-1}.
\]

**Approximate Singular Values** If $\sigma_i \approx \gamma_i$, $1 \leq i \leq k^* \leq k$,

$\sigma_{k^*}^2 / (\sigma_{k^*}^2 + \alpha^2) >> \sigma_i^2 / (\sigma_i^2 + \alpha^2) \approx 0$, $i > k^*$,

\[
\text{Tr}(A(\alpha)) \approx \text{Tr}(B_{k^*}(\alpha)) + \sum_{i=k^*+1}^{m^*} \sigma_i^2 (\sigma_i^2 + \alpha^2)^{-1} \approx \text{Tr}(B_{k^*}(\alpha)).
\]

If $k^*$ approx numerical rank $A$, $\zeta_k^{\text{opt}} \approx \alpha^{\text{opt}}$ for $K_{k^*}(A^T A, A^T b)$. 
$A \in \mathcal{R}^{m \times n}$, target rank $k$, oversampling parameter $p$, $k + p = kp \ll m$. Power factor $q$. Compute $A \approx A_k = U_k \Sigma_k V_k^T$, $U_k \in \mathcal{R}^{m \times k}$, $\Sigma_k \in \mathcal{R}^{k \times k}$, $V_k \in \mathcal{R}^{n \times k}$.

1: Generate a Gaussian random matrix $\Omega \in \mathcal{R}^{n \times kp}$.
2: Compute $Y = A\Omega \in \mathcal{R}^{m \times kp}$. $Y = \text{orth}(Y)$
3: If $q > 0$ repeat $q$ times \{ $Y = A(A^TY)$, $Y = \text{orth}(Y)$ \}. Power
4: Form $B = Y^TA \in \mathcal{R}^{kp \times n}$. ($Q = Y$)
5: Economy SVD $B = U_B \Sigma_B V_B^T$, $U_B \in \mathcal{R}^{kp \times kp}$, $V_B \in \mathcal{R}^{k \times k}$
6: $U_k = QU_B(:, 1:k)$, $V_k = V_B(:, 1:k)$, $\Sigma_k = \Sigma_B(1:k, 1:k)$
7: Then $A_k = U_k \Sigma_k V_k^T$

Provides an approximate TSVD

Convergence and regularization parameter estimation?

Comparison to LSQR?
Theory for the approximation of the TSVD by the RSVD

**Theorem (Proto)**  
*Target rank* \( k \geq 2, \) **oversampling** \( p \geq 2, \) \( k + p \leq \min\{m, n\}. \)

\[
E(\|A - QQ^T A\|) \leq \left[ 1 + \frac{4\sqrt{k + p}}{p - 1} \cdot \sqrt{\min\{m, n\}} \right] \sigma_{k+1}
\]

**Theorem (Power Iteration to force singular values to 0)**  
*Exponent* \( q, \) **target** \( k \) **singular values,** \( 2 \leq k \leq 0.5 \min\{m, n\}. \)

*Then rank-2\( k \) factorization \( U\Sigma V^T : \)

\[
E(\|A - U\Sigma V^T\|) \leq \left[ 1 + 4 \sqrt{\frac{2 \min\{m, n\}}{k - 1}} \right]^{1/(2q+1)} \sigma_{k+1}
\]

Or truncate to get a rank \( k \) approximation so that

\[
E(\|A - U\Sigma_{(k)} V^T\|) \leq \left( 1 + \left[ 1 + 4 \sqrt{\frac{2\min\{m, n\}}{k - 1}} \right]^{1/(2q+1)} \right) \sigma_{k+1}
\]
Approximation The RSVD gives a near best rank \( k \) approximation on the average.

UPRE Because the RSVD yields approximate TSVD - the UPRE regularization convergence should apply. (No new analysis required as for the LSQR)
Contrasting the RSVD and LSQR

**RSVD** RSVD with standard oversampling. \((p = k)\)

**RSVDQ** RSVD with power iteration and \(q = 2\). \((p = k)\)

**LSQR** Standard LSQR

**LSQRO** Oversample in the LSQR using \(p = k\) to find \(B_{p+k}\) and its SVD. Use relevant \(k\) components of the SVD as for the RSVD.

**Aims**
1. Compare running times
2. Compare spectral approximation
3. Compare regularization estimation
Contrasting the RSVD and LSQR

Singular Values: RSVD - LSQR - LSQRO - RSVDQ

Relative Error in Singular Values 2–norm error

The LSQRO errors are small as compared to RSVD and RSVDQ
Contrasting the RSVD and LSQR: II

Running time and approximation error

- Running time is high for LSQRO.
- Approximation error for LSQR is high but not for LSQRO.
- Singular values are better approximated by RSVDQ and LSQRO
- Relative errors in solutions high for LSQR

$\alpha_k$ converges with $k$ when singular space approximated well: RSVD, LSQRO, RSVDQ
Example Solutions for Phillips (Trivial)

Parameter $k$ increasing [4, 12, 20, 28, 36, 44, 52, 60]
Observations comparing RSVD and LSQR

- RSVD spectrum does not inherit ill-conditioning of $A$
- LSQR spectrum does inherit ill-conditioning - unless oversampling is applied
- Dominant spectrum of RSVD is less accurate than LSQR (with or without oversampling)
- Degree of ill-conditioning is relevant. - convergence
- RSVD is far cheaper to find dominant spectrum
- Accuracy of SVD does not imply accuracy of solution - regularization required - problem ill-posed.
- LSQR generates a solution with information from entire spectrum and thus is unstable
- Oversampling in LSQR eliminates the issues with LSQR semiconvergence with increasing $k$ (the instability).

LSQRO as effective as RSVD but expensive
Restoration of Grain noise level $\sigma^2 = .0001$:

Relative Errors decrease with TSVD approximation.

Regularization parameter converges as $k$ increases.
Restored Regularized Solutions noise level $\sigma^2 = .0001$

RSVD - RSVDP - LSQR - LSQRPO

$k = 1200$

RSVDREG
RSVDQREG
LSQRREG
LSQROREG

$k = 2000$

RSVDREG
RSVDQREG
LSQRREG
LSQROREG
Iteratively Reweighted Regularization [LK83]

\[ \|Ax - b\|^2 + \alpha^2 \|L^{(\ell)}(x^{(\ell)} - x^{(\ell-1)})\|^2 \]

Minimum Support Stabilizer Regularization operator \(L^{(\ell)}\).

\[
(L^{(\ell)})_{ii} = \left(\left(x_i^{(\ell-1)} - x_i^{(\ell-2)}\right)^2 + \beta^2\right)^{-1/2} \quad \beta > 0
\]

Parameter \(\beta\) ensures \(L^{(\ell)}\) invertible

Invertibility use \((L^{(\ell)})^{-1}\) as right preconditioner for \(A\)

\[
(L^{(\ell)})_{ii}^{-1} = \left(\left(x_i^{(\ell-1)} - x_i^{(\ell-2)}\right)^2 + \beta^2\right)^{1/2} \quad \beta > 0
\]

Initialization \(L^{(0)} = I, \ x^{(0)} = x_0\). (might be 0)

Reduced System When \(\beta = 0\) and \(x_i^{(\ell-1)} = x_i^{(\ell-2)}\) remove column \(i\), \(\hat{A}\) is \(\hat{A}L^{-1}\) with columns removed.

Update Equation Solve \(\hat{A}\hat{y} \approx R = b - Ax^{(\ell-1)}\). With correct indexing set \(y_i = \hat{y}_i\) if updated, else \(y_i = 0\).

\[ x^{(\ell)} = x^{(\ell-1)} + y \]

Cost of \(L^{(\ell)}\) is minimal
Undersampled Gravity data $m = 5500$, $n = 66000$ $\beta^2 = 1e - 9$, $k = 1000$
Conclusions

**UPRE** converges for the TSVD

**UPRE** therefore converges for the RSVD

**UPRE** converges for LSQR with oversampling

$\zeta_k^{\text{opt}}, \alpha^{\text{opt}}$ related across levels in all cases

**Regularization** Find the optimal parameter for reduced subspace and apply for larger number of terms.

**Underdetermined** problems - also possible.

**Extensions** approaches apply in other contexts - eg generalized Tikhonov / iterative reweighting (L1)

**RSVD or LSQR** Results suggest advantages of the RSVD - speed!

**LSQR** Run with oversampling to avoid issues of semi-convergence.
Thank you

Questions
Some key references

Julianne Chung, James G Nagy, and DIANNE P O'Leary.
A weighted GCV method for Lanczos hybrid regularization.

Percy Deift, James Demmel, Luen-Chau Li, and Carlos Tomei.
The bidiagonal singular value decomposition and hamiltonian mechanics.

N. Halko, P. G. Martinsson, and J. A. Tropp.
Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix
decompositions.

Z. Jia.
The Regularization Theory of the Krylov Iterative Solvers LSQR, CGLS, LSMR and CGME For Linear
Discrete Ill-Posed Problems.

B. J. Last and K. Kubik.
Compact gravity inversion.

R. A. Renaut, S. Vatankhah, and V. E. Ardestani.
Hybrid and iteratively reweighted regularization by unbiased predictive risk and weighted gcv for projected
systems.