Least squares, the singular value decomposition, and linear inverse problems

Marcel Oliver

Last revised July 20, 2022

1 Simple least squares

The *simple least squares* is the following special case of a *linear regression* problem: Find the equation of a line which is "closest" to a given set of points in the plane. More precisely, given tuples of real numbers $(x_1, y_1), \dots, (x_n, y_n)$, find numbers a and b such that

$$f(a,b) \equiv \sum_{i=1}^n (a\,x_i + b - y_i)^2 \tag{1}$$

is minimal.

As f is a smooth function defined for all $(a,b) \in \mathbb{R}^2$, calculus tells us that it can only have a minimum provided its partial derivatives vanish. I.e.,

$$0 = \frac{\partial f}{\partial b} = 2 \sum_{i=1}^n a\,x_i + b - y_i \tag{2a}$$

and

$$0 = \frac{\partial f}{\partial a} = 2 \sum_{i=1}^n x_i\,(a\,x_i + b - y_i)\,. \tag{2b}$$

Defining the *mean* of the $\{x_i\}$ and $\{y_i\}$ as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \qquad \text{and} \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i\,, \tag{3}$$

equation (2a) can be written

$$b = \bar{y} - a\,\bar{x}\,. \tag{4a}$$

Substituting into (2b) and solving for a, we find

$$a = \frac{\displaystyle\sum_{i=1}^n x_i\, y_i - n\,\bar{x}\,\bar{y}}{\displaystyle\sum_{i=1}^n x_i^2 - n\,\bar{x}^2} = \frac{\displaystyle\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^n (x_i - \bar{x})^2} \,. \tag{4b}$$

Thus, we can first compute a via (4b) and then b via (4a).

Two remarks are in order. First, vanishing of the partial derivatives is, in general, only a necessary condition for the existence of a minimum. Here, however, it is not difficult to show that $f(a,b) \to \infty$ as $a, b \to \infty$ (in any way) provided there are at least two distinct data points. Thus, since the necessary condition yields a unique solution, this solution must correspond to a minimum of f.

Second, the minimization of the Euclidean norm of the error as opposed to some other error measure is certainly convenient as it yields a smooth and even linear relationship between measured y-values and estimates for the coefficients. However, it is often also preferred for statistical reasons: When the errors are assumed Gaussian, the least-squares fit equals the so-called maximum-likelihood estimator. When the errors are all from the same probability distribution, least-squares is the so-called best unbiased linear estimator. We will not dwell on these issues for the time being.

2 Calculus on normed vector spaces

In the following, we give a simple framework for doing calculus on arbitrary normed[1] vector spaces.[2] A typical first year course on multivariate calcu-

[1] A normed vector space $\mathcal{V}$ is a vector space together with a map $\|\cdot\| \colon \mathcal{V} \to [0,\infty)$ such that

(i) $\|v\| = 0$ if and only if $v = 0$,

(ii) $\|\lambda v\| = |\lambda|\,\|v\|$ for all $\lambda \in \mathbb{R}$ and $v \in \mathcal{V}$, and

(iii) $\|v + w\| \leq \|v\| + \|w\|$ for all $v, w \in \mathcal{V}$.

The last property is called the *triangle inequality.* When $\mathcal{V} \in \mathbb{R}^n$, we take without further discussion the *Euclidean Norm* defined via

$$\|\boldsymbol{v}\|^2 = \sum_{i=1}^n v_i^2 = \boldsymbol{v}^T \boldsymbol{v}$$

for $\boldsymbol{v} = (v_1, \dots, v_n)^T \in \mathbb{R}^n$.

[2] For simplicity, we only consider vector spaces over $\mathbb{R}$. In case that the space is infinite dimensional, we would typically assume completeness, i.e., that we have a Banach space.

lus typically considers the vector space $\mathbb{R}^n$ only. However, looking at more general vector spaces is in some sense easier, and naturally includes differentiation with respect to matrices as well as the calculus of variations, both of which occur frequently in applications. In this basic introduction, we make no attempt to give a rigorous definition of differentiability nor do we discuss any of the analytical subtleties of doing calculus on infinite dimensional vector spaces such as function spaces.

Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces and $f \colon \mathcal{V} \to \mathcal{W}$ a continuous function. Central to the understanding of differentiation in this setting is the notion of *directional derivative*. Namely, fix $v \in \mathcal{V}$ and choose an arbitrary $\delta v \in \mathcal{V}$.[3] We define the directional derivative as the instantaneous rate of change of f when changing v by one unit of δv. We write

$$\mathrm{d}f(v)[\delta v] = \lim_{h \to 0} \frac{f(v + h\,\delta v) - f(v)}{h} = \frac{\mathrm{d}f(v + h\,\delta v)}{\mathrm{d}h}\bigg|_{h=0} . \tag{5}$$

If $\mathrm{d}f$ is continuous on $\mathcal{V} \times \mathcal{V}$, for each $v \in \mathcal{V}$, the map $\mathrm{d}f(v)[\,\cdot\,] \colon \mathcal{V} \to \mathcal{W}$, the called the *total derivative*, is linear.[4] I.e., for fixed $v \in \mathcal{V}$,

$$\mathrm{d}f(v)[\lambda \delta v] = \lambda\, \mathrm{d}f(v)[\delta v]\,, \tag{6a}$$

$$\mathrm{d}f(v)[\delta v + \delta w] = \mathrm{d}f(v)[\delta v] + \mathrm{d}f(v)[\delta w] \tag{6b}$$

for all $\delta v, \delta w \in \mathcal{V}$ and $\lambda \in \mathbb{R}$.

Practical computation The following trick often simplifies the computation of the derivative. Writing

$$\delta f \equiv \mathrm{d}f(v)[\delta v]\,, \tag{7}$$

the δ symbol formally behaves like differentiation,[5] so we can apply the product rule, chain rule, etc. as usual.

Example 1. On $\mathcal{V} = \mathbb{R}^n$, fix an $A \in M(n \times n)$ and define

$$f(\boldsymbol{v}) = \boldsymbol{v}^T A \boldsymbol{v}\,. \tag{8}$$

[3] We must issue an important warning: In typical generalizations, v and δv are taken from different sets. For example, $\mathcal{V}$ may be an affine space $a + \mathcal{U}$. Then the "admissible variations" δv must be taken from $\mathcal{U}$. Or $\mathcal{V}$ may be a manifold, then δv is an element from its tangent space $T\mathcal{V}$.

[4] In the context of infinite dimensional Banach spaces, the directional derivative is called the Gâteaux derivative and the total derivative leads to so-called Fréchet derivative. For Gâteaux differentiability to imply Fréchet differentiability, one needs, moreover, continuity of $\mathrm{d}f \colon \mathcal{V} \to \mathcal{L}(\mathcal{V}, \mathcal{W})$.

[5] In fact, with only a change of language we could call this a differential.

Then, using the product rule,

$$\delta f = (\delta \boldsymbol{v})^T A \boldsymbol{v} + \boldsymbol{v}^T A \delta \boldsymbol{v} = \boldsymbol{v}^T A^T \delta \boldsymbol{v} + \boldsymbol{v}^T A \delta \boldsymbol{v} \tag{9}$$

so that $\mathrm{d}f(\boldsymbol{v})[\delta \boldsymbol{v}] = \boldsymbol{v}^T (A + A^T) \delta \boldsymbol{v}$ or $\mathrm{d}f(\boldsymbol{v}) = \boldsymbol{v}^T (A + A^T)$. You probably know that the derivative of a real-valued function on $\mathbb{R}^n$ can be computed via its Jacobian, which here is the $1 \times n$ matrix

$$\mathrm{d}f(\boldsymbol{v}) = (\partial_1 f, \ldots, \partial_n f) \, . \tag{10}$$

Sure enough, using this formula, you'll find the same answer, but with a lot more pain. Try it!

Example 2. Let

$$\mathcal{V} = \{\phi \in C^2([0,1]) \colon \phi(0) = \phi(1) = 0\} \tag{11}$$

and fix $g \in C([0,1])$. Define

$$f(\phi) = \int_0^1 \bigl(\tfrac{1}{2} \, (\phi'(x)^2 + g(x) \, \phi(x) \bigr) \, \mathrm{d}x \, . \tag{12}$$

Then

$$\delta f = \int_0^1 \bigl(\phi'(x) \, \delta\phi'(x) + g(x) \, \delta\phi(x) \bigr) \, \mathrm{d}x = \int_0^1 (-\phi''(x) + g(x)) \, \delta\phi(x) \, \mathrm{d}x \tag{13}$$

where, in the last equality, we have used integration by parts noting that the boundary terms vanish due to the definition of $\mathcal{V}$. Thus, (13) is an expression for $\mathrm{d}f(\phi)[\delta\phi]$; we note that there is no elementary way to separate out $\delta\phi$ in this expression.

Extrema We consider maps $f \colon \mathcal{V} \to \mathbb{R}$, as in the examples above, and ask for conditions under which such maps assume local minima or maxima. Throughout, we assume that $f \in C^1(\mathcal{V}, \mathbb{R})$, i.e., f is at least once continuously differentiable. In this case, $\mathrm{d}f(v)$ is a linear functional (i.e., a linear map from $\mathcal{V}$ to $\mathbb{R}$) called the *gradient* of f. We say that f has a local maximum at $v \in \mathcal{V}$ if there is $\varepsilon > 0$ such that $f(v) \geq f(w)$ for all $\|v - w\| < \varepsilon$; f has a local minimum if $-f$ has a local maximum. Then we have the following necessary condition for the existence of a local extreme point:

> f has a local minimum or maximum at $v \in V$ only if V is a *stationary point* of f, i.e., if $\mathrm{d}f(v) = 0$.

The proof is a direct extension of the proof of the "first derivative test" from single-variable calculus and shall not be given here.

Note that if f is only defined on a domain $D \subsetneq V$, then it may assume a local or global extremum on the boundary (with obvious modification to the definition of the local extremum). This case may have to be considered separately. The technique of Lagrange multipliers below can be useful if the boundary can be described by a constraint.

Lagrange Multipliers Sometimes we need to find extreme values of f where v is subject to additional constraints. In the simplest case, assume that the constraint is given by the equation $g(v) = 0$ for some $g \in C^1(\mathcal{V}, \mathbb{R})$. We have the following necessary condition.

> f has a local minimum or maximum on the set $\{v \in V : g(v) = 0\}$ only if there exists $\lambda \in \mathbb{R}$ such that
>
> $$\mathrm{d}f(v) = \lambda \, \mathrm{d}g(v) \,. \tag{14}$$

We will not give a proof, but shall give the following geometric motivation. For the purposes of discussion, suppose that $\mathcal{V} = \mathbb{R}^n$. First, we notice that the gradient of a function is a vector pointing in the direction of steepest ascent.[6] Thus, if we vary $\boldsymbol{v}$ in the constraint set $\{v \in V : g(v) = 0\}$ in the direction $\delta \boldsymbol{v} = \mathrm{d}g(\boldsymbol{v})^T$ leave the constraint set. On the other hand, any variation $\delta \boldsymbol{v} \perp \mathrm{d}g(\boldsymbol{v})^T$ is tangent to the constraint set and therefore permitted. Thus, when "probing" f for a local extremum, we can allow f to ascent or decent in the "forbidden" direction $\mathrm{d}g(\boldsymbol{v})^T$. On the other hand, when f's steepest ascent direction $\mathrm{d}f(\boldsymbol{v})$ aligns with the "forbidden" direction $\mathrm{d}f(\boldsymbol{v})$, then the instantaneous rate of change of f is zero in any direction tangent to the constraint set. This alignment is represented by the Lagrange multiplier condition (14).

This argument also points out how to generalize to K constraints: Each constraint function $g_1, \dots, g_K$ induces a "forbidden" direction $\mathrm{d}g_i(\boldsymbol{v})^T$ which must be excluded from the "search" when probing f for an extremum. Thus, the ascent direction of f may lie in the "forbidden" space spanned by all

[6] To see this, recall that in Euclidean geometry,

$$\boldsymbol{u}^T \boldsymbol{w} = \|\boldsymbol{u}\| \, \|\boldsymbol{w}\| \, \cos \angle(\boldsymbol{u}, \boldsymbol{v}) \,.$$

Thus, if $\boldsymbol{u} = \mathrm{d}f(\boldsymbol{v})$, the instantaneous rate of change in the direction of $\delta \boldsymbol{v}$ can be written $\boldsymbol{u}^T \delta \boldsymbol{v}$ and is therefore maximal among all vectors of length $\|\boldsymbol{u}\|$ if $\delta \boldsymbol{v} = \boldsymbol{u}$.

the $\mathrm{d}g_i(\boldsymbol{v})^T$. Thus, a necessary condition for a local extremum under K constraints is the existence of a Lagrange multiplier $\boldsymbol{\lambda} \in \mathbb{R}^K$ such that

$$\mathrm{d}f(v) = \boldsymbol{\lambda}^T \mathrm{d}\boldsymbol{g}(v)\,. \tag{15}$$

3 General linear least squares

Recall that in Section 1, the model for our data is the linear equation

$$y = a\,x + b\,. \tag{16}$$

We generalize as follows. Assume $\boldsymbol{x} \in \mathbb{R}^m$ with $m \in \mathbb{N}$ while we still take $y \in \mathbb{R}$. (The situation discussed here easily generalizes to vector-valued "output data" y by considering each of its components separately.) So our data is again a set of tuples $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)$. A *linear model* for our data is a linear combination with unknown coefficients of a fixed finite set of possibly nonlinear "ansatz functions" $h_i \colon \mathbb{R}^m \to \mathbb{R}$ for $i = 1, \dots, k$. I.e., we seek to find $\boldsymbol{v} \in \mathbb{R}^k$ such that

$$y = \boldsymbol{h}(\boldsymbol{x})^T \boldsymbol{v} \tag{17}$$

with $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \dots, h_k(\boldsymbol{x}))^T$ fits the data with minimal mean-square error

$$f(\boldsymbol{v}) = \sum_{i=1}^n (\boldsymbol{h}(\boldsymbol{x}_i)^T \boldsymbol{v} - y_i)^2\,. \tag{18}$$

Setting

$$H = \begin{pmatrix} h_1(\boldsymbol{x}_1) & \dots & h_k(\boldsymbol{x}_1) \\ \vdots & & \vdots \\ h_1(\boldsymbol{x}_n) & \dots & h_k(\boldsymbol{x}_n) \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \tag{19}$$

we can interpret this problem as the minimization of the so-called residual $H\boldsymbol{v} - \boldsymbol{y}$ corresponding to the potentially inconsistent linear system $H\boldsymbol{v} = \boldsymbol{y}$. (In the following, nothing depends on the fact that these equations come from a general linear fitting problem; we are generally asking for the least-square solution of some inconsistent system of linear equations.) In matrix notation, we ask for the minimum of

$$f(\boldsymbol{v}) = \|H\boldsymbol{v} - \boldsymbol{y}\|^2 = (H\boldsymbol{v} - \boldsymbol{y})^T (H\boldsymbol{v} - \boldsymbol{y})\,. \tag{20}$$

The necessary condition for a minimum of f is

$$0 = \delta f = (H\delta \boldsymbol{v})^T(H\boldsymbol{v} - \boldsymbol{y}) + (H\boldsymbol{v} - \boldsymbol{y})^T H\delta\boldsymbol{v} = 2\,(H\boldsymbol{v} - \boldsymbol{y})^T H\delta\boldsymbol{v} \tag{21}$$

for any $\delta\boldsymbol{v} \in \mathbb{R}^k$, which implies $(H\boldsymbol{v})^T H = \boldsymbol{y}^T H$ or

$$(\boldsymbol{y} - H\boldsymbol{v})^T H = \mathbf{0}\,. \tag{22}$$

Geometric interpretation In the language of Linear Algebra, equation (22) can be seen as an orthogonality condition, namely[7]

$$\boldsymbol{y} - H\boldsymbol{v} \perp \operatorname{Range} H\,, \tag{23}$$

i.e., the process of minimizing $f(\boldsymbol{v})$ yields a decomposition

$$\boldsymbol{y} = \underbrace{\boldsymbol{y} - H\boldsymbol{v}}_{\perp \operatorname{Range} H} + \underbrace{H\boldsymbol{v}}_{\in \operatorname{Range} H}\,. \tag{24}$$

Thus, the point $H\boldsymbol{v}$ is the point on the hyperplane $\operatorname{Range} H$ of closest Euclidean distance to $\boldsymbol{y}$.

Solvability In practical terms, we solve (22) by solving the linear system

$$H^T H\boldsymbol{v} = H^T\boldsymbol{y}\,, \tag{25}$$

called the *normal equations*. It is an easy exercise in Linear Algebra to verify that $\operatorname{Range} H^T = \operatorname{Range} H^T H$, therefore (25) has a solution $\boldsymbol{v}$ for any $\boldsymbol{y} \in \mathbb{R}^n$ and can be solved directly. From the computational perspective, this is usually most efficient. However, when the $k \times k$ matrix $H^T H$ is invertible, we may also write

$$\boldsymbol{v} = (H^T H)^{-1} H^T \boldsymbol{y}\,. \tag{26}$$

Example 3. We can recover the result of Section 1 by setting $m = 1$, $k = 2$, $h_1(x) = x$, $h_2(x) = 1$, and $\boldsymbol{v} = (a, b)^T$. Equation (26) is thus a compact way of writing (4b) and (4a).

[7] We define, as usual,

$$\operatorname{Range} H = \{H\boldsymbol{u} \colon \boldsymbol{u} \in \mathbb{R}^k\}$$

and

$$\operatorname{Ker} H = \{\boldsymbol{u} \in \mathbb{R}^k \colon H\boldsymbol{u} = \mathbf{0}\}\,.$$

Example 4. Suppose we would like to fit a polynomial of degree $k-1$ to a set of data points taken at integer points $x_1 = 0, \dots, x_n = n-1$. If we choose as ansatz functions $h_1(x) = 1, h_2(x) = x, \dots, h_k(x) = x^{k-1}$, H is the *Vandermonde matrix*

$$H = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2^2 & \cdots & 2^{k-1} \\ \vdots & & & & \vdots \\ 1 & n-1 & (n-1)^2 & \cdots & (n-1)^{k-1} \end{pmatrix} . \tag{27}$$

It is poorly conditioned for even moderate values of k; as a result, polynomial least squares with a degree higher than about 3–6 are rarely a good idea.

When $H^T H$ is not invertible, i.e., when $\operatorname{Ker} H \neq \{\mathbf{0}\}$, there is linear dependence in the data points or in the ansatz functions. In this case, the effective number of independent observations is less than k. In this case, we may wish to compute a least-norm solution of (25), discussed next.

4 Least norm solutions

When a system of linear equations

$$A\boldsymbol{v} = \boldsymbol{b}\,, \tag{28}$$

where $A \in M(n \times k)$ is consistent, but does not have a unique solution (i.e., if $\boldsymbol{b} \in \operatorname{Range} A$ and $\operatorname{Ker} A \neq \{\mathbf{0}\}$), we may ask for a solution $\boldsymbol{v}$ which has the least norm amongst the family of solutions. (In the context of the general linear least squares problem of Section 3, $A = H^H$ and $\boldsymbol{b} = H^T\boldsymbol{y}$.)

Geometric interpretation If $\boldsymbol{w}$ is any solution of $A\boldsymbol{w} = \boldsymbol{b}$, then it can be decomposed as

$$\boldsymbol{w} = \underbrace{\boldsymbol{w} - \boldsymbol{v}}_{\in \operatorname{Ker} A} + \underbrace{\boldsymbol{v}}_{\perp \operatorname{Ker} A}\,, \tag{29}$$

where $\boldsymbol{v}$ is the minimum norm solution. Thus, $\|\boldsymbol{v}\|$ is the distance of $\boldsymbol{w}$ to the hyperplane $\operatorname{Ker} A$. This decomposition is the domain-side analog to the range-side decomposition (24) for least-square solutions.

Formulation as constrained minimization The least norm problem can be formulated as the following constrained minimization problem.

Find $\boldsymbol{v}$ minimizing $f(\boldsymbol{v}) = \boldsymbol{v}^T\boldsymbol{v}$ subject to $\boldsymbol{g}(\boldsymbol{v}) = A\boldsymbol{v} - \boldsymbol{b} = 0$.

Since $\mathrm{d}f(\boldsymbol{v}) = 2\boldsymbol{v}^T$ and $\mathrm{d}\boldsymbol{g}(\boldsymbol{v}) = A$, the method of Lagrange multipliers (15) yields the necessary condition $2\boldsymbol{v}^T = \boldsymbol{\lambda}^T A$ or, setting $\boldsymbol{\mu} = -2\boldsymbol{\lambda}$,

$$I\boldsymbol{v} + A^T\boldsymbol{\mu} = \boldsymbol{0}\,. \tag{30}$$

Together with the constraint $A\boldsymbol{v} = \boldsymbol{b}$, we can write the necessary condition in block-matrix form,

$$\begin{pmatrix} I & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{v} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{b} \end{pmatrix}\,. \tag{31}$$

It is easy to check that this system is consistent as long as (28) is consistent. Moreover,

$$\operatorname{Ker} \begin{pmatrix} I & A^T \\ A & 0 \end{pmatrix} = \left\{ \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{u} \end{pmatrix} \in \mathbb{R}^{k+n} \colon \boldsymbol{u} \in \operatorname{Ker} A^T \right\}. \tag{32}$$

In other words, the solution of (31) exists and, if not unique, the non-uniqueness affects only the Lagrange multiplier $\boldsymbol{\mu}$; the least-norm vector $\boldsymbol{v}$ is uniquely determined.

5 Singular value decomposition (SVD)

Recall from your linear algebra class that if $S \in M(k \times k)$ is a symmetric matrix, it has an orthogonal diagonalization. More precisely, there exists an orthogonal matrix[8] $V \in M(k \times k)$ such that

$$D = V^T S V \tag{33}$$

is diagonal. The orthogonal diagonalizability of symmetric matrices can be used to construct a decomposition of *any* matrix into a product of orthogonal and diagonal matrices. We proceed as follows.

Let $A \in M(n \times k)$. Without loss of generality, we may assume that $n \geq k$. Then $S = A^T A \in M(k \times k)$ is symmetric and has an orthogonal diagonalization of the form (33). Moreover, S is positive semidefinite.[9] By

[8] A matrix $V \in M(k \times k)$ is orthogonal if $V^{-1} = V^T$. This is equivalent to $V^T V = I$ (the columns of V are an orthogonal basis of the column space) which is equivalent to $VV^T = I$ (the rows of V are an orthogonal basis of the row space of V).

[9] A matrix $S \in M(k \times k)$ is positive definite if $\boldsymbol{v}^T S\boldsymbol{v} > 0$ for all nonzero $\boldsymbol{v} \in \mathbb{R}^k$; it is positive semi-definite if $\boldsymbol{v}^T S\boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^k$. It is easy to see that all eigenvalues of a positive definite matrix must be positive; all eigenvalues of a positive semidefinite matrix must be nonnegative.

permuting the columns of V, we can always achieve that the diagonal entries d_{ii} of D appear in decreasing order so that, in particular, $d_{11}, \dots, d_{mm} > 0$ and $d_{m+1,m+1}, \dots, d_{kk} = 0$ for some $m \le k$. We now define

$$\Sigma_0 = \sqrt{D} \equiv \begin{pmatrix} \sqrt{d_{11}} & 0 & & & \cdots & 0 \\ 0 & \ddots & & & & \vdots \\ & & \sqrt{d_{mm}} & & & \\ & & & 0 & & \\ \vdots & & & & \ddots & \\ 0 & \cdots & & & & 0 \end{pmatrix} \tag{34}$$

and[10]

$$\Sigma_0^\dagger = \begin{pmatrix} d_{11}^{-1/2} & 0 & & & \cdots & 0 \\ 0 & \ddots & & & & \vdots \\ & & d_{mm}^{-1/2} & & & \\ & & & 0 & & \\ \vdots & & & & \ddots & \\ 0 & \cdots & & & & 0 \end{pmatrix} . \tag{35}$$

and

$$U_0 = AV\Sigma_0^\dagger \, . \tag{36}$$

First, we note that the first m columns of U_0 are an orthonormal set in the column space, as

$$U_0^T U_0 = \Sigma_0^\dagger V^T A^T A V \Sigma_0^\dagger = \Sigma_0^\dagger V^T S V \Sigma_0^\dagger = \Sigma_0^\dagger D \Sigma_0^\dagger = \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} . \tag{37}$$

Second,

$$U_0 \Sigma_0 V^T = AV\Sigma_0^\dagger \Sigma_0 V^T = A \, . \tag{38}$$

The last identity is obvious if A is nonsingular so that $m = k$ and $\Sigma_0^\dagger \Sigma_0 = I$. In general, note that $A^T A = VDV^T$ implies

$$\operatorname{Ker} A^\perp = \operatorname{Span}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_m\} \tag{39a}$$

[10] $\Sigma_0^\dagger$ is called the *pseudo-inverse* of Σ_0. It is characterized by

$$\Sigma_0^\dagger \Sigma_0 = \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} ,$$

where I_m denotes the $m \times m$ identity matrix.

and

$$\operatorname{Ker} A = \operatorname{Span}\{\boldsymbol{v}_{m+1}, \dots, \boldsymbol{v}_k\}\,, \tag{39b}$$

where $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ denote the columns of V. The last step of (38) is then easily verified by restricting to $\operatorname{Ker} A$ and $\operatorname{Ker} A^\perp$, respectively. Equation (38) is referred to as the *restricted singular value decomposition* of A.

Now let $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$ denote the first m columns of U_0. Note that (37) shows that

$$\operatorname{Range} A = \operatorname{Span}\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_m\}\,. \tag{39c}$$

It is often convenient to extend the orthonormal set $\boldsymbol{u}_1, \dots, \boldsymbol{u}_m$ to an orthonormal basis $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ of $\mathbb{R}^n$, so that

$$\operatorname{Range} A^\perp = \operatorname{Span}\{\boldsymbol{u}_{m+1}, \dots, \boldsymbol{u}_n\}\,. \tag{39d}$$

Then, setting

$$U = \begin{pmatrix} | & & | \\ \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_n \\ | & & | \end{pmatrix} \in M(n \times n) \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_0 \\ 0 \end{pmatrix} \in M(n \times k)\,, \tag{40}$$

we can write

$$A = U \Sigma V^T \tag{41}$$

This is the *singular value decomposition* of A; the diagonal entries of Σ, denoted $\sigma_1, \dots, \sigma_k$, are the *singular values* of A, the vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k$ are the *right singular vectors*, and the vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_n$ are the *left singular vectors.*

Variational characterization Recall, from Homework 1, that the principal unit eigenvector of a symmetric matrix $S \in M(k \times k)$ maximizes the function

$$f(\boldsymbol{v}) = \boldsymbol{v}^T S \boldsymbol{v} \tag{42}$$

among all unit vectors $\boldsymbol{v} \in \mathbb{R}^k$. The singular value decomposition of a matrix $A \in M(n \times k)$ has a similar variational characterization. Consider the function

$$f(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^T A \boldsymbol{v}\,. \tag{43}$$

Then the maximum of f over the set of unit vectors $\boldsymbol{u} \in \mathbb{R}^n$ and unit vectors $\boldsymbol{v} \in \mathbb{R}^k$ is the largest singular value of A; vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ at which the maximum is attained can be taken as the first left singular vector $\boldsymbol{u}_1$ and the first right singular vector $\boldsymbol{v}_1$, respectively, in the notation of (39).

To establish our claim, we proceed as follows. Our task can be formulated as an optimization problem for f on $\mathbb{R}^{n+k}$ with the two scalar constraints

$$g_1(\boldsymbol{u},\boldsymbol{v}) \equiv \boldsymbol{u}^T\boldsymbol{u} - 1 = 0 \quad \text{and} \quad g_2(\boldsymbol{u},\boldsymbol{v}) \equiv \boldsymbol{v}^T\boldsymbol{v} - 1 = 0\,. \tag{44}$$

Clearly,

$$\mathrm{d}f(\boldsymbol{u},\boldsymbol{v})[\delta\boldsymbol{u},\delta\boldsymbol{v}] = \boldsymbol{v}^T A^T \delta\boldsymbol{u} + \boldsymbol{u}^T A \delta\boldsymbol{v}\,, \tag{45a}$$

$$\mathrm{d}g_1(\boldsymbol{u},\boldsymbol{v})[\delta\boldsymbol{u},\delta\boldsymbol{v}] = 2\,\boldsymbol{u}^T\delta\boldsymbol{u}\,, \tag{45b}$$

and

$$\mathrm{d}g_2(\boldsymbol{u},\boldsymbol{v})[\delta\boldsymbol{u},\delta\boldsymbol{v}] = 2\,\boldsymbol{v}^T\delta\boldsymbol{v}\,. \tag{45c}$$

The necessary condition for a local maximum is given by the Lagrange multiplier principle (15) which reads, using (45) and rearranging terms,

$$\bigl(\boldsymbol{v}^T A^T \delta\boldsymbol{u} - 2\,\lambda_1\,\boldsymbol{u}^T\delta\boldsymbol{u}\bigr) + \bigl(\boldsymbol{u}^T A \delta\boldsymbol{v} - 2\,\lambda_2\,\boldsymbol{v}^T\delta\boldsymbol{v}\bigr) = 0 \tag{46}$$

which, since $\delta\boldsymbol{u}$ and $\delta\boldsymbol{v}$ can be chosen independently, implies

$$\boldsymbol{v}^T A^T = 2\,\lambda_1\,\boldsymbol{u}^T \quad \text{and} \quad \boldsymbol{u}^T A = 2\,\lambda_2\,\boldsymbol{v}^T\,. \tag{47}$$

Right-multiplying the first equality with $\boldsymbol{u}$ and the second equality with $\boldsymbol{v}$, we conclude that

$$\boldsymbol{u}^T A \boldsymbol{v} = 2\,\lambda_1 = 2\,\lambda_2 \equiv \sigma\,. \tag{48}$$

The existence of a solution is clear as we are maximizing a smooth function over a compact constraint set; hence the necessary condition must be satisfied somewhere. Denote some such solution triple $\boldsymbol{u}_1$, $\boldsymbol{v}_1$, and σ_1.[11] Now, restricting $\boldsymbol{u}$ and $\boldsymbol{v}$ to the respective orthogonal complements of the vectors already established, we can iterate this argument k times to obtain a full set of "singular values" $\sigma_1,\dots,\sigma_k$, "right singular vectors" $\boldsymbol{v}_1,\dots,\boldsymbol{v}_k$ and, completing the orthogonal set to a basis, "left singular vectors" $\boldsymbol{u}_1,\dots,\boldsymbol{u}_n$.

Two claims remain to be verified. First, that the result of the above construction is indeed indeed a singular value decomposition. Second, that the largest singular value is indeed the global (not just some local) maximum of f as claimed. The verification of both statements is not difficult and shall be left as an exercise.

[11] We are not claiming uniqueness. In fact, it can be shown that the solution is unique up to a choice of sign if the singular values of A are nondegenerate; in the degenerate case, it is clear that we cannot expect uniqueness.

Matrix norms Given any norm $\|\cdot\|$ on $\mathbb{R}^n$, we can define the norm of a matrix $A = M(n \times k)$ as the smallest number $\|A\|$ such that

$$\|A\boldsymbol{x}\| \leq \|A\| \, \|\boldsymbol{x}\| \tag{49}$$

holds true. In other words,

$$\|A\| = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \,. \tag{50}$$

This definition generalizes to operators between Banach spaces and is often referred to as the *operator norm.* Note that, if the coefficients of A are denoted a_{ij},

$$\|A\|_F^2 = \sum_{\substack{i \in 1,\ldots,n \\ j \in 1,\ldots,k}} |a_{ij}|^2 \tag{51}$$

defines a norm, called the *Frobenius norm*, which is is not an operator norm.

Consider, in particular, the operator norm induced by the Euclidean norm on $\mathbb{R}^n$, which we generally assume in these notes. Then $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x}$, so that, using the singular value decomposition,

$$\begin{aligned} \|A\|^2 &= \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^T V \Sigma^T U^T U \Sigma V^T \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} \\ &= \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}^T V \Sigma^T \Sigma V^T \boldsymbol{x}}{\boldsymbol{x}^T V V^T \boldsymbol{x}} \\ &= \max_{\boldsymbol{y} \neq \boldsymbol{0}} \frac{\boldsymbol{y}^T \Sigma^T \Sigma \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{y}} \\ &= \sigma_1^2 \end{aligned} \tag{52}$$

In other words, the matrix norm induced by the Euclidean metric equals the largest singular value of the matrix.

In contrast, it is not difficult to show that

$$\|A\|_F^2 = \sum_{i=1}^{k} \sigma_i^2 \,. \tag{53}$$

Condition numbers Suppose, for simplicity, that $A \in M(n \times n)$ is a non-singular matrix. We would like to characterize the stability of the solution $\boldsymbol{x}$ to the system of linear equations

$$A\boldsymbol{x} = \boldsymbol{b} \tag{54}$$

under perturbation of the right hand side. By linearity, $\Delta \boldsymbol{b} = A \Delta \boldsymbol{x}$, where $\Delta \boldsymbol{b}$ denotes an *absolute* change of the vector $\boldsymbol{b}$, and $\Delta \boldsymbol{x}$ the corresponding absolute change in $\boldsymbol{x}$. Thus, by (49), we obtain $\|\Delta \boldsymbol{x}\| \leq \|A^{-1}\| \, \|\Delta \boldsymbol{b}\|$ and $\|\boldsymbol{b}\| \leq \|A\| \, \|\boldsymbol{x}\|$, so that

$$\frac{\|\Delta \boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\| \, \|A^{-1}\| \, \frac{\|\Delta \boldsymbol{b}\|}{\|\boldsymbol{b}\|} \, . \tag{55}$$

The worst-case amplification factor for *relative* differences,

$$\kappa(A) = \|A\| \, \|A^{-1}\| \, , \tag{56}$$

is called the *condition number* of A. Due to (52),

$$\kappa(A) = \frac{\sigma_1}{\sigma_n} \, . \tag{57}$$

This notion of condition number generalizes naturally to non-square matrices A of full rank. When the condition number is large, the solution $\boldsymbol{x}$ becomes very sensitive to error (e.g. measurement error) in the data $\boldsymbol{b}$; we speak of an *ill-conditioned* problem.

6 Approximate solution of ill-conditioned linear equations

Let us consider the following prototypical version of an ill-conditioned system of linear equations. Our presentation closely follows [2]. Let $\boldsymbol{x} \in \mathbb{R}^n$ represent the true pixel values of a, for simplicity, one-dimensional image. The image is taken by a device which, due to lens imperfections, say, records a blurred set of pixel values $\boldsymbol{b} \in \mathbb{R}^n$. We assume a simple linear model for the device imperfections, namely that

$$b_i = \sum_{j=1}^{n} a_{i-j} \, x_j \, , \tag{58}$$

where $b_1, \ldots, b_n$ and $x_1, \ldots, x_n$ denote the components of $\boldsymbol{b}$ and $\boldsymbol{x}$, respectively. This type of expression is referred to as a *discrete convolution* with *kernel* a_k; typical is a bell-like shape of its graph, e.g.,

$$a_k = c \, \mathrm{e}^{-\gamma k^2} \tag{59}$$

for some given constants c and γ. Equation (58) can be written as a system of linear equations

$$\boldsymbol{b} = A \boldsymbol{x} \tag{60}$$

with $n \times n$ matrix

$$A = \begin{pmatrix} a_0 & a_{-1} & \cdots & & \\ a_1 & \ddots & \ddots & & \\ \vdots & \ddots & & \ddots & \vdots \\ & & \ddots & \ddots & a_{-1} \\ & & \cdots & a_1 & a_0 \end{pmatrix}. \tag{61}$$

Matrices of this type are known as *Toeplitz matrices*. They are typically very ill-conditioned[12], so that, even in the case that A is non-singular, the direct solution of (60) may be practically meaningless.

Regularization The ill conditioning of (60) can be understood in terms of the singular values of A as follows. Suppose $\boldsymbol{x}_{\text{true}}$ refers to the true pixel values of the image; due to measurement error, the observed value $\boldsymbol{b}$ can be written

$$\boldsymbol{b} = A\boldsymbol{x}_{\text{true}} + \boldsymbol{\eta}\,, \tag{62}$$

where $\boldsymbol{\eta}$ is the measurement error vector with norm $\|\boldsymbol{\eta}\| > 0$. For simplicity, we suppose that $A \in M(n \times n)$ is nonsingular. Thus, when attempting to compute the pixel values by direct inversion of A, we obtain

$$\boldsymbol{x}_{\text{naive}} = A^{-1}\boldsymbol{b} = \boldsymbol{x}_{\text{true}} + V\Sigma^{-1}U^T\boldsymbol{\eta}\,. \tag{63}$$

The ill-conditioning manifests itself in the occurrence of small singular values for A, so that Σ^{-1} in (63) has small denominators—some components of $\boldsymbol{\eta}$ with respect to the orthonormal basis $\{\boldsymbol{v}_i\}$ are being multiplied with large gains. As a result, we cannot expect $\boldsymbol{x}_{\text{naive}}$ and $\boldsymbol{x}_{\text{true}}$ to be close.

Since the problem is due to small denominators, let us define a filtered solution

$$\boldsymbol{x}_\alpha = V\Sigma_\alpha^{-1}U^T\boldsymbol{b} \tag{64}$$

[12] In fact, (58) can be interpreted as a discrete approximation of the integral equation

$$b(x) = \int_a^b k(x-y)\,f(y)\,\mathrm{d}y$$

where, with a smooth kernel function k, the right hand side defines a compact, hence non-invertible operator on a Hilbert space. Such problem are referred to as *ill-posed*. Direct discretization of an ill-posed continuum problem always leads to an ill-conditioned finite dimensional system. In this sense, the ill-conditioning in our example is a generic phenomenon.

where Σ_α^{-1} is the filtered inverse of Σ, defined by

$$\Sigma_\alpha^{-1} = \operatorname{diag}\{w_\alpha(\sigma_i^2)\sigma_i^{-1} \colon i = 1, \ldots, n\} \tag{65}$$

with $w(\sigma^2)$ converging to zero at least linearly as $\sigma \to 0$ to avoid unbounded singular values of the filtered inverse in this limit. Among the many possible choices for w_α, the following are particularly natural.

Truncation The simplest choice is the cut-off filter

$$w_\alpha(\sigma^2) = \begin{cases} 1 & \text{if } \sigma^2 \geq \alpha \\ 0 & \text{if } \sigma^2 < \alpha \,. \end{cases} \tag{66}$$

Tychonov regularization With the choice

$$w_\alpha(\sigma^2) = \frac{\sigma^2}{\sigma^2 + \alpha}\,, \tag{67}$$

we obtain

$$w_\alpha(\sigma^2)\sigma^{-1} = \frac{\sigma}{\sigma^2 + \alpha} \tag{68}$$

or

$$\Sigma_\alpha^{-1} = (\Sigma^2 + \alpha I)^{-1}\Sigma\,, \tag{69}$$

thereby providing a smooth transition between the filtered and unfiltered regime. Substituting this expression into (64), we obtain

$$\begin{aligned} \boldsymbol{x}_\alpha &= V(\Sigma^2 + \alpha I)^{-1}\Sigma U^T\boldsymbol{b} \\ &= V(\Sigma^2 + \alpha I)^{-1}V^T V\Sigma U^T\boldsymbol{b} \\ &= (V(\Sigma^2 + \alpha I)V^T)^{-1}V\Sigma U^T\boldsymbol{b} \\ &= (A^T A + \alpha I)^{-1}A^T\boldsymbol{b}\,. \end{aligned} \tag{70}$$

Note that this expression formally resembles the normal equations (26) of the linear least squares problem, suggesting a close relationship between the two problems. This expectation will be substantiated later when we describe a variational formulation of the Tychonov regularization.

Sparse matrices In many applications, the matrix A is sparse, or is well-approximated by a sparse matrix. Thus, it is important that the regularization can be implemented such that it can be implemented without losing sparsity.[13] The truncation filter will certainly not work in this setting, as it depends on an explicit singular value decomposition which, even if only the singular values up to the filter scale α were ever computed, would amount to doing computations with full matrices.

The Tychonov regularization in the form (70), on the other hand, is equivalent to solving the linear system

$$(A^T A + \alpha I)\boldsymbol{x}_\alpha = A^T \boldsymbol{b} \,, \tag{71}$$

cf. the normal equations (25) for the least-square problem. If A is sparse, the system matrix is also sparse and iterative solvers can be used to solve the system in approximately $O(N)$ time, where N is the number of non-zero entries of A.

Error analysis We provide a deterministic error analysis, i.e., we assume that $\boldsymbol{\eta}$ is a fixed measurement error vector of known magnitude. We define

$$\begin{aligned}\boldsymbol{e}_\alpha = \boldsymbol{x}_{\text{true}} - \boldsymbol{x}_\alpha &= \boldsymbol{x}_{\text{true}} - V\Sigma_\alpha^{-1}U^T\boldsymbol{b} \\ &= VIV^T\boldsymbol{x}_{\text{true}} - V\Sigma_\alpha^{-1}U^T(A\boldsymbol{x}_{\text{true}} + \boldsymbol{\eta}) \\ &= V(I - \Sigma_\alpha^{-1}\Sigma)V^T\boldsymbol{x}_{\text{true}} - V\Sigma_\alpha^{-1}U^T\boldsymbol{\eta} \\ &\equiv \boldsymbol{e}_{\text{trunc}} + \boldsymbol{e}_{\text{noise}} \end{aligned} \tag{72}$$

Note that, in contrast to the noise term in (63), the expression for $\boldsymbol{e}_{\text{noise}}$ involves a filtered inverse, thus has bounded gains. We expect that $\|\boldsymbol{e}_{\text{noise}}\|$ decreases with increasing filter parameter α, while the truncation error $\|\boldsymbol{e}_{\text{trunc}}\|$ increases; thus, there should be an optimal filter parameter α. This is borne out by the following estimation.

First, due to (49) and the fact that the Euclidean matrix norm of an orthogonal matrix equals 1,

$$\|\boldsymbol{e}_{\text{noise}}\| \le \|V\| \, \|\Sigma_\alpha^{-1}\| \, \|U^T\| \, \|\boldsymbol{\eta}\| \le \max_{\sigma>0} \frac{w_\alpha(\sigma^2)}{\sigma} \, \|\boldsymbol{\eta}\| \le \frac{1}{2\sqrt{\alpha}} \, \|\boldsymbol{\eta}\| \,. \tag{73}$$

The last inequality above can be verified independently for the truncation filter and for the Tychonov regularization.

Second, we seek to derive estimates on the truncation error. Here we are faced with a fundamental dilemma. In order to conclude that truncation

[13] A linear operator A stored as 64-bit floating point numbers acting on a grayscale image of size 512^2 stored as a full matrix will require $8 \times 512^2 \times 512^2 = 512\,\text{GiB}$ of memory, substantially more than available as main memory on a desktop-class computer.

error—the error committed by filtering—is small, we need to ensure that the information which gets removed by the filter is is some sense unimportant for the true solution. In other words, we need to provide additional *a priori* information on the problem.[14] In a standard analysis of this problem, one assumes the *range condition*[15]

$$\boldsymbol{x}_{\text{true}} = A^T \boldsymbol{z} \tag{74}$$

such that $\|\boldsymbol{z}\|$ is "reasonable," meaning of the same order of magnitude as $\|\boldsymbol{x}_{\text{true}}\|$. It is important to realize that this condition, from the mathematical point of view, is arbitrary. Rather, it represents additional knowledge about the specific problem providing additional restrictions on the the class of vectors in which we seek a solution. As such, it must be justified on physical grounds which necessarily lie outside of the analysis presented here.

Proceeding with our estimation, we obtain, inserting the range condition (74) into the expression for $\boldsymbol{e}_{\text{trunc}}$,

$$\begin{aligned} \|\boldsymbol{e}_{\text{trunc}}\|^2 = \boldsymbol{e}_{\text{trunc}}^T \, \boldsymbol{e}_{\text{trunc}} &= \boldsymbol{z}^T AV(I - \Sigma_\alpha^{-1}\Sigma)V^T V(I - \Sigma_\alpha^{-1}\Sigma)V^T A^T \boldsymbol{z} \\ &= \boldsymbol{z}^T U \, (I - \Sigma_\alpha^{-1}\Sigma)^2 \, \Sigma^2 \, U^T \boldsymbol{z} \\ &\leq \max_{\sigma>0} (1 - w_\alpha(\sigma^2))^2 \, \sigma^2 \, \|\boldsymbol{z}\|^2 \\ &\leq \frac{\alpha}{4} \, \|\boldsymbol{z}\|^2 \, . \end{aligned} \tag{75}$$

Inserting the bounds for $\boldsymbol{e}_{\text{trunc}}$ and $\boldsymbol{e}_{\text{noise}}$ into (72), we obtain

$$\|\boldsymbol{e}_\alpha\| \leq \frac{\sqrt{\alpha}}{2} \, \|\boldsymbol{z}\| + \frac{1}{2\sqrt{\alpha}} \, \|\boldsymbol{\eta}\| \, . \tag{76}$$

It is easy to show that the right hand side is minimized when the two term balance, i.e., $\alpha = \|\boldsymbol{\eta}\|/\|\boldsymbol{z}\|$, so that

$$\|\boldsymbol{e}_\alpha\| = \sqrt{\|\boldsymbol{\eta}\| \, \|\boldsymbol{z}\|} \, . \tag{77}$$

We call this an *a priori* estimate for the regularization parameter since knowledge of the true solution $\boldsymbol{x}_{\text{true}}$ is required. Next, we discuss a near-optimal choice for the regularization parameter knowing only the data.

[14] In fact, in the absence of further information, we can still conclude, directly from the definition of the w_α considered, that $\|\boldsymbol{e}_{\text{trunc}}\| \to 0$ as $\alpha \to 0$. This conclusion, however, is practically meaningless because we have no control about the rate of convergence, so there is no way to ensure that we are doing substantially better than the naive, direct inversion (63).

[15] In an infinite dimensional Hilbert space setting, the range condition can be stated concisely as

$$\boldsymbol{x}_{\text{true}} \in \operatorname{Range} A^T \, .$$

There, this condition is nontrivial.

Morozov discrepancy principle In contrast to what was done in (72), we can split the error into two components as follows,[16]

$$\boldsymbol{e}_\alpha^* \equiv A(\boldsymbol{x}_{\text{true}} - \boldsymbol{x}_\alpha) = (\boldsymbol{b} - A\boldsymbol{x}_\alpha) + \boldsymbol{\eta} \equiv \boldsymbol{e}_{\text{trunc}}^* + \boldsymbol{e}_{\text{noise}}^* \,. \tag{78}$$

Since $\boldsymbol{e}_{\text{trunc}}^*$ now only depends on the measured data $\boldsymbol{b}$, we can compute an *a posteriori* estimate of α by balancing the norms of $\boldsymbol{e}_{\text{trunc}}^*$ and $\boldsymbol{e}_{\text{noise}}^*$ rather than those of $\boldsymbol{e}_{\text{trunc}}$ and $\boldsymbol{e}_{\text{noise}}$ as in (76). Notice that

$$\boldsymbol{e}_{\text{trunc}}^* = U(I - \Sigma\Sigma_\alpha^{-1})U^T\boldsymbol{b} \tag{79}$$

so that

$$\|\boldsymbol{e}_{\text{trunc}}^*\|^2 = \boldsymbol{b}^T U \, (I - \Sigma\Sigma_\alpha^{-1})^2 \, U^T \boldsymbol{b} \to \begin{cases} 0 & \text{as } \alpha \to 0 \\ \boldsymbol{b}^T\boldsymbol{b} & \text{as } \alpha \to \infty \,; \end{cases} \tag{80}$$

the left hand expression is, moreover, smooth and monotonic. Thus, whenever $\|\boldsymbol{b}\| \geq \|\boldsymbol{\eta}\|$, the intermediate value theorem guarantees the existence of an α_{opt} such that $\|\boldsymbol{e}_{\text{trunc}}^*(\alpha_{\text{opt}})\| = \|\boldsymbol{\eta}\|$. This regularization parameter selection criterion is called the *Morozov discrepancy principle.*

Error analysis of the Morozov principle We now ask the question how the error $\boldsymbol{e}_\alpha$ behaves under the Morozov discrepancy principle. This question is only easy to answer in the case of the Tychonov regularization where $\boldsymbol{x}_\alpha$ is known to minimize the function

$$f_\alpha(\boldsymbol{x}) = \|A\boldsymbol{x} - \boldsymbol{b}\|^2 + \alpha \, \|\boldsymbol{x}\|^2 \tag{81}$$

over all $\boldsymbol{x} \in \mathbb{R}^n$; see Homework 2. Thus, in particular,

$$f_\alpha(\boldsymbol{x}_\alpha) \leq f_\alpha(\boldsymbol{x}_{\text{true}}) \tag{82}$$

which we can write

$$\|A\boldsymbol{x}_\alpha - \boldsymbol{b}\|^2 + \alpha \, \|\boldsymbol{x}_\alpha\|^2 \leq \|\boldsymbol{\eta}\|^2 + \alpha \, \|\boldsymbol{x}_{\text{true}}\|^2 \,. \tag{83}$$

When α is chosen according to the Morozov principle, $\|A\boldsymbol{x}_\alpha - \boldsymbol{b}\| = \|\boldsymbol{\eta}\|$, so that the first terms on left and right of this inequality drop out and we conclude that

$$\|\boldsymbol{x}_\alpha\| \leq \|\boldsymbol{x}_{\text{true}}\| \,. \tag{84}$$

[16] The definition $\boldsymbol{e}_\alpha^* = A\boldsymbol{e}_\alpha$ seems arbitrary, but is motivated by the need to avoid any inversion of A as this computation generally cannot be performed in a stable manner.

Then,

$$
\begin{aligned}
\|\boldsymbol{e}_\alpha\|^2 &= \|\boldsymbol{x}_{\text{true}}\|^2 - 2\,\boldsymbol{x}_\alpha^T\boldsymbol{x}_{\text{true}} + \|\boldsymbol{x}_\alpha\|^2 \\
&\le 2\,\|\boldsymbol{x}_{\text{true}}\|^2 - 2\,\boldsymbol{x}_\alpha^T\boldsymbol{x}_{\text{true}} = 2\,\boldsymbol{e}_\alpha^T\boldsymbol{x}_{\text{true}}\,. \qquad (85)
\end{aligned}
$$

As before, we assume the range condition with $\boldsymbol{x}_{\text{true}} = A^T\boldsymbol{z}$, so that

$$
\|\boldsymbol{e}_\alpha\|^2 \le 2\,\boldsymbol{e}_\alpha^T A^T\boldsymbol{z} = 2\,(\boldsymbol{e}_\alpha^*)^T\boldsymbol{z} \le 2\,\|\boldsymbol{e}_\alpha^*\|\,\|\boldsymbol{z}\|\,. \qquad (86)
$$

When α is chosen according to the Morozov principle, $\|\boldsymbol{e}_\alpha^*\| = 2\,\|\boldsymbol{\eta}\|$. Inserting this estimate back into (86) and taking the square root, we obtain

$$
\|\boldsymbol{e}_\alpha\| = 2\,\sqrt{\|\boldsymbol{\eta}\|\,\|\boldsymbol{z}\|} \qquad (87)
$$

which coincides with the *a priori* error estimate (77) up to a factor 2.

Final Remarks The range condition is crucial for getting good reconstructions. In the case when A represents a discrete convolution or the inverse of a discrete differential operator, the range condition qualitatively expresses that $\boldsymbol{x}_{\text{true}}$ consists of samples of a smooth function on a grid. A concise discussion is beyond the scope of these notes, but the consequences of this statement are easily visible in simple computational experiments. In particular, the filtered inverse will reconstruct or reproduce step functions only poorly—the edges will get eroded.

In the variational characterization (81), the term $\alpha\,\|\boldsymbol{x}\|^2$ can be seen as a "penalty term" which might be replaced with other types of penalties, not all of which can be described in terms of the singular value decomposition. One important example is to replace $\|\boldsymbol{x}\|^2$ by the so-called *total variation.* The total variation penalizes oscillatory behavior while allowing jumps. The book by Vogel [2] gives a more complete, yet relatively accessible introduction to theory and computational methods for inverse problems.

References

[1] Å. Björck, "Numerical Methods for Least Squares Problems," SIAM, Philadelphia, 1996.

[2] C. Vogel, "Computational Methods for Inverse Problems," SIAM, Philadelphia, 2002.