Continuous Regularization Methods

Aditya Ganeshan

Indian Institute of Technology, Roorkee

December 13, 2016

Table of Content

Introduction

A a-posteriori parameter choice rule that is always optimal!
(Upto the qualification μ_0 of the regularization method.)

From the earlier estimate of error, we have,

$$\left\|x_\alpha^\delta - x^\dagger\right\| \leqslant \left\|x_\alpha - x]^\dagger\right\| + \delta\sqrt{CG_\alpha} \tag{1}$$

Squaring it we get,

$$\left\|x_\alpha^\delta - x^\dagger\right\|^2 \leqslant 2(\left\|r_\alpha(T^*T)x^\dagger\right\|^2 + C\delta^2 G_\alpha) \tag{2}$$

Minimizing the right hand side should certainly yield a good choice for α. Hence,

$$\frac{\partial}{\partial\alpha}(\left\|r_\alpha(T^*T)x^\dagger\right\|^2 + C\delta^2 G_\alpha) = 0 \tag{3}$$

Lemma 4.11

In addition to the assumptions made abut g_α in Theorem 4.1, we assume that g_α and g_α are continuously differentiable with respect to α and that $\frac{\partial G}{\partial \alpha} \neq 0$. Then for $\omega \in D(T^\dagger)$,

$$\frac{\partial}{\partial \alpha}(\|r_\alpha(T^*T)\omega\|^2) = -\frac{\partial G_\alpha}{\partial \alpha} f(\alpha, \omega) \tag{4}$$

where,

$$f(\alpha, \omega) = 2\left(\frac{\partial G_\alpha}{\partial \alpha}\right)^{-1} \left\langle \frac{\partial g_\alpha}{\partial \alpha}(TT^*) r_\alpha(TT^*) Q\omega, Q\omega \right\rangle \tag{5}$$

Remark

Due to Lemma 4.11, The condition (5) is equivalent to ,

$$f(\alpha, y) = C\delta^2 \tag{6}$$

Assumption

Let g_α fulfill the assumptions made in Theorem 4.1, and assume, in addition, that $\alpha \to g_\alpha$ and $\alpha \to g_\alpha$ is continuously differentiable, that there exists a constant $K > 0$ such that,

$$\left| \frac{\partial g_\alpha}{\partial \alpha}(\lambda) \left(\frac{\partial G_\alpha}{\partial \alpha} \right)^{-1} \right| \leqslant K \tag{7}$$

hold for all $\alpha > 0, \lambda \geqslant 0$, and that the function,

$$\alpha \Rightarrow \left(\frac{\partial G_\alpha}{\partial \alpha} \right)^{-1} \frac{\partial g_\alpha}{\partial \alpha}(\lambda) r_\alpha(\lambda) \tag{8}$$

is strictly increasing for all $\lambda > 0$

Lemma 4.12

Under the previous assumption, the function f is, for $Q\omega \neq 0$, continuous and strictly increasing in α. Furthermore,

$$\lim_{\alpha \to 0} f(\alpha, y) = 0$$

and

$$\lim_{\alpha \to +\infty} f(\alpha, y) = h(\omega)$$

with,

$$h(\omega) = 2 \int \lim_{\alpha \to +\infty} \left[\left(\frac{\partial G_\alpha}{\partial \alpha} \right)^{-1} \frac{\partial g_\alpha}{\partial \alpha}(\lambda) r_\alpha(\lambda) \right] d\|F_\lambda Q\omega\|^2 \qquad (9)$$

Where $\{F_\lambda\}$ is the spectral family generated by TT^*.

Proposition (4.13)

Under the given assumptions, for any $\delta > 0$ *and* $y^\delta \in Y$ *with* $Qy^\delta \neq 0$, *there is a unique* $\alpha = \alpha(\delta, y^\delta) > 0$, *such that*

$$f(\alpha, y^\delta) = \tau\delta^2 \tag{10}$$

holds, provided that

$$\tau \in (0, h(y^\delta)\delta^{-2}) \tag{11}$$

Proposition (4.14)

With the assumptions and notation of proposition 4.13, let

$$L = 2sup\{\left|\left[\frac{\partial G_\alpha}{\partial \alpha}\right]^{-1}\frac{\partial g_\alpha}{\partial \alpha}(\lambda) r_\alpha(\lambda)\right| \mid \alpha > 0, \lambda \geqslant 0\} \tag{12}$$

and let $\tau > L, Qy \neq 0$. *For* $\delta > 0$, *let* $y^\delta \in Y$ *be such that* $\|y^\delta - y\| \leqslant \delta$. *Furthermore,let*

$$\tau_1 = (\sqrt{tau} - \sqrt{L})^2, \quad \tau_2 = (\sqrt{tau} + \sqrt{L})^2 \tag{13}$$

Proposition (Continued)

If $Qy^\delta \neq 0$, (11) *holds, and* $\alpha(\delta, y^\delta)$ *is defined via* (10), *then there is a* $\tau = \tau(\delta, y^\delta) \in [\tau_1, \tau_2]$, *such that*

$$f(\alpha(\delta, y^\delta), y) = \tau(\delta, y^\delta)\delta^2 \tag{14}$$

holds. (11) *and* $Qy^\delta \neq 0$ *hold especially if,*

$$\delta^2 < \frac{h(y)}{\tau_2} \tag{15}$$

Theorem (4.15)

Let the Assumption hold, let $y \in R(T)$, $y \neq 0$, *let* $\tau > L$, *and let, for* $\delta > 0, y^\delta \in Y$ *fulfill* $\|y^\delta - y\|$, *and* (11).*Then if* $\alpha = \alpha(\delta, y^\delta)$ *is determined as the unique solution of* (10), *there is a constant* η *such that with* x^δ, x^δ_α *as defined earlier,*

$$\left\| x^\delta_{\alpha(\delta, y^\delta)} - T^\dagger y \right\| \leqslant \eta \; inf\{\|x_\alpha - x^\dagger\| + \delta\sqrt{CG_\alpha} \mid \alpha > 0\} \qquad (16)$$

Corollary (4.16)

Let the assumptions of Corollary 4.4, and the above assumptions be fulfilled. Then, the regularization method $(R_\alpha, \alpha(\delta, y^\delta))$ is of optimal order in $X_{\mu,\rho}$ and convergent for all $y \in R(T)$.

Remark

Thus, the parameter choice strategy (10) is of optimal order in $X_{\mu,\rho}$ for all μ for which,

$$\omega_\mu \sim \alpha^\mu \quad \textit{for} \quad 0 < \mu \leqslant \mu_0$$

Hence, this parameter choice strategy gives optimal order up to the qualification μ_0.

The condition (11) and (14), together with requirement that $\tau > L$, can be intercepted as "signal to noise ratio conditions".

If $Qy^\delta = y^\delta$ then the parameter choice rule that we have analyzed in Theorem 4.15 defines $\alpha(\delta, y^\delta)$ as a solution of the non-linear equation

$$\eta_\alpha = \left\langle y^\delta, s_\alpha(TT^*)y^\delta \right\rangle = \tau\delta^2, \tag{17}$$

where

$$s_\alpha(\lambda) = \left(\frac{\partial G_\alpha}{\partial \alpha}\right)^{-1} \frac{\partial g_\alpha}{\partial \alpha}(\lambda) r_\alpha(\lambda) \tag{18}$$

and

$$\tau > \gamma = \sup_{\alpha,\lambda>0} |s_\alpha(\lambda)| \tag{19}$$

The discrepancy principle is defined in a similar form, namely

$$\alpha(\delta, y^{\delta}) = \sup\{\alpha > 0 \mid \eta_{\alpha} \leqslant \tau\delta^2\} \tag{20}$$

where η_{α} is given by (17).

Theorem (4.17)

Let the assumptions of Theorem 4.3 hold with ,

$$\omega_\mu(\alpha) = c\alpha^\mu, \quad 0 < \mu \leqslant \mu_0 < \infty$$

and some $c > 0$, and let G_α as defined earlier, fulfill

$$G_\alpha = O(\frac{1}{\alpha}). \tag{21}$$

Theorem (Continued)

Let $0 \leqslant v \leqslant \mu_0$, *and* $\{s_\alpha\}$ *be a family of positive and piece wise continuous functions with,*

$$s_\alpha(\lambda) \sim \left(\frac{\alpha}{\lambda+\alpha}\right)^{2v+1} \tag{22}$$

Furthermore, assume that $s_\alpha(\lambda)$ *is continuous from the left with respect to* α. *Then the parameter choice* (20), *with* η_α *defined by* (17), *and* τ *subject to* (19) *is order-optimal in* $X_{\mu,\rho}$ *for all* $\rho > 0$ *and* $0 < \mu \leqslant v$.

Table of Content

The previous a-posteriori parameter choice rules all depend in one way or the other on the computed approximation and the data error level δ.
In real world, such noise level information is not always available.
The worst-case bound may be an overestimation of the error level, while other measures such as standard deviations might underestimate the error level.

Both estimate might lead to significant loss in accuracy when used in parameter choice rule.

Hence, often it is necessary to consider alternative parameter choice rules that avoid knowledge of the noise levels.
Such heuristic parameter choice rules are called *error free*.

Most heuristic rules for error free parameter choice rules are based on some kind of error estimation.We moniter the norm of the residual.

$$y^\delta - Tx_\alpha^\delta = r_\alpha(TT^*)x^\dagger + r_\alpha(TT^*)(y^\delta - y) \tag{23}$$

How does this relate to the actual error?

$$x^\dagger - x_\alpha^\delta = r_\alpha(TT^*)x^\dagger + g_\alpha(T^*T)T^*(y - y^\delta) \tag{24}$$

The above representation is split into two components, corresponding to the exact solution and the data-error, respectively.

From earlier results we have, provided $x^\dagger \in X_{\mu,\rho}$,

$$\left\| r_\alpha(T^*T)T^*x^\dagger \right\| < \rho\omega_{\mu+\frac{1}{2}}(\alpha) \tag{25}$$

$$\left\| r_\alpha(T^*T)x^\dagger \right\| < \rho\omega_{\mu+}(\alpha) \tag{26}$$

Assuming, the qualification of the method is $\mu_0 = \infty$, and

$$\omega_\mu(\alpha) = O(\alpha^\mu), \quad 0 \leqslant \mu_0 < \infty \tag{27}$$

Thus, we conclude that the upper bound for $\left\| r_\alpha(T^*T)x^\dagger \right\|$ decays faster by a factor of $\sqrt{\alpha}$ as $\alpha \to 0$. This extra factor is *independent* of the actual value of μ. Thus, we can expect,

$$\frac{1}{\alpha} \left\| r_\alpha(T^*T)T^*x^\dagger \right\| \sim \left\| r_\alpha(T^*T)x^\dagger \right\| \tag{28}$$

Now, we compare the propagated noise component of the rescaled residual and the approximation error.

$$\frac{1}{\alpha} r_\alpha(T^*T)(y^\delta - y) \quad vs. \quad g_\alpha(T^*T)T^*(y^\delta - y)$$

The following figure shows the Showalter's method with $\alpha = 0.05$.

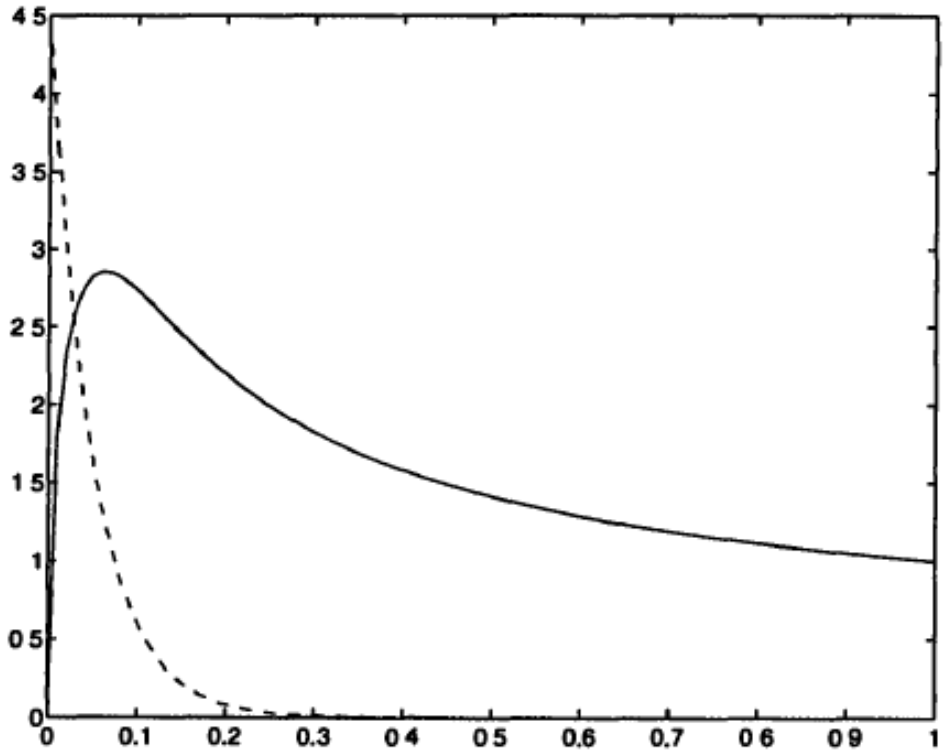

Figure 4.1: The functions $\lambda^{\frac{1}{2}} g_\alpha(\lambda)$ (solid) and $\alpha^{-\frac{1}{2}} r_\alpha(\lambda)$ (dashed).

We conclude that although the two functions are quite different for larger values of λ, they are pretty similar for $\lambda \sim \alpha$.

Moreover, if $G_\alpha = O(\alpha_{-1})$ then the overall maxima of the two function are both of the order $1/\sqrt{\alpha}$ and are both attained for $\lambda \sim \alpha$.

Hence we take $\left\|y^\delta - Tx_\alpha^\delta\right\| / \sqrt{\alpha}$, and pick the regularization parameter as the one for which

$$\frac{1}{\sqrt{\alpha}} \left\|y^\delta - Tx_\alpha^\delta\right\| \to min. \tag{29}$$

Theorem (4.18)

Assume that the functions $\{r_\alpha\}$ and $\{g_\alpha\}$ satisfy the assumption of Theorem 4.1 and that

$$\omega_\mu = O(\alpha_\mu), \quad 0 < \mu \leqslant \mu_0, \quad G_\alpha = O\left(\frac{1}{\alpha}\right) \tag{30}$$

Without loss of generality, let $\delta = \|y^\delta - y\| < \|y\|$, and assume that $x^\dagger \in X_{\mu,\rho}$, for some $0 < \mu \leqslant \mu_0 - 1/2$. If the global minimizer $\alpha(y^\delta)$ of (29) over $\alpha \in (0, \|T\|^2]$ exists, and if

$$\delta_* = \left\| y^\delta - T x^\delta_{\alpha(y^\delta)} \right\| \neq 0,$$

then,

$$\left\| x^\dagger - x^\delta_{\alpha(y^\delta)} \right\| \leqslant c\left(1 + \frac{\delta}{\delta}\right) max\{\delta, \delta_*\}^{\frac{2\mu}{2\mu+1}} \rho^{\frac{1}{2\mu+1}} \tag{31}$$

Remark

This theorem states that the approximation obtained from this heuristic parameter choice rule is order-optimal, provided δ_* is of the same order as δ.

The knowledge of δ_* can be used for a a-posteriori check of this parameter choice rule: if $\delta_* << \delta$, we have to be very cautious about the chosen parameter; if $\delta_* >> \delta$,the situation is not critical and the magnitude of δ_* determines the error.

Remark

Many difficulties originate from looking at the residual norm, $\|y^\delta - Tx_\alpha^\delta\|$. Hence instead of this, some other quantity is employed in the parameter choice rule.The following theorem shows an example.

Theorem (4.19)

Under the assumptions of Theorem 4.10, let η_α *be defined by*

$$\eta_\alpha = \left\langle y^\delta, s_\alpha(TT^*)y^\delta \right\rangle \tag{32}$$

and let $\alpha(y^\delta)$ *be the minimizer of* η_α/α *in* $(0, \|T\|^2]$. *Additionally, let* $\delta = \|y^\delta - y\| < \|y\|$ *and* $x^\dagger \in X_{\mu,\rho}$, *for some* $0 < \mu \leqslant \mu_0$. *If*

$$\delta_* = \eta^{\frac{1}{2}}_{\alpha(y^\delta)} > 0 \tag{33}$$

then

$$\left\| x^\dagger - x^\delta_{\alpha(y^\delta)} \right\| \leqslant c\left(1 + \frac{\delta}{\delta_*}\right) max\{\delta, \delta_*\}^{\frac{2\mu}{2\mu+1}} \rho^{\frac{1}{2\mu+1}}$$

Another Heuristic Method

Instead of looking at the usual residual norm as in Theorem 4.18, we can also look at the norm of the residual with respect to the normal equations,

$$T^*(y^\delta - Tx_\alpha^\delta) = r_\alpha(TT^*)TT^*x^\dagger + r_\alpha(TT^*)T^*(y^\delta - y) \tag{34}$$

Similar considerations as above suggest that under the same assumptions as before

$$\frac{1}{\alpha}\left\|T^*(y^\delta - Tx_\alpha^\delta)\right\| \tag{35}$$

may be a useful error estimate if $\mu_0 = \infty$

Another Heuristic Method

The graph for the respective functions is given is figure 2.

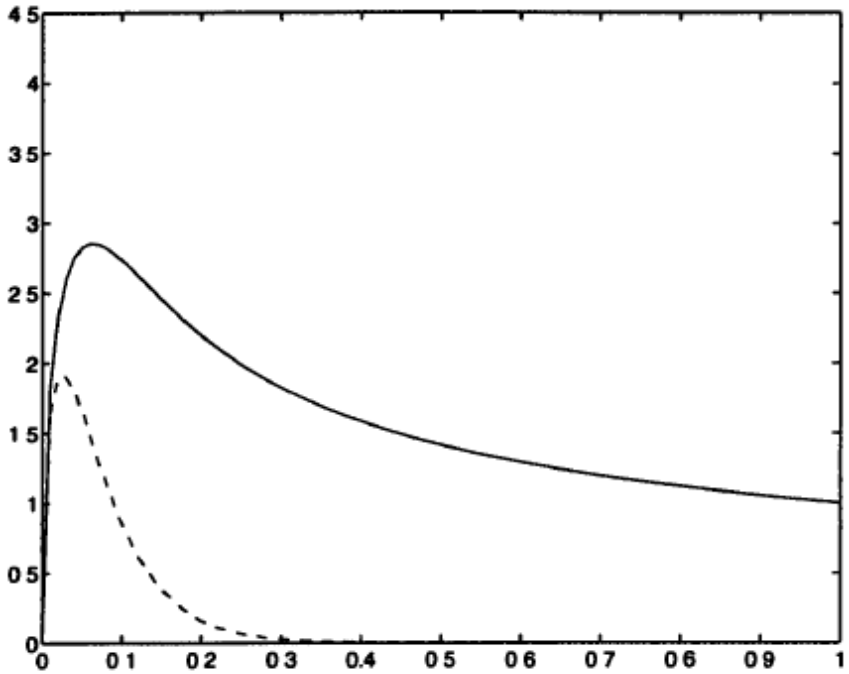

Figure 4.2: The functions $\lambda^{\frac{1}{2}} g_\alpha(\lambda)$ (solid) and $\frac{\lambda^{\frac{1}{2}}}{\alpha} r_\alpha(\lambda)$ (dashed).

Generalized Cross-validation

It applies to the problem where $T = K$ is an operator into a finite dimensional data space.
we shall write $y = [y_1, y_2 \ldots y_m]^T$ for the entire data vector, and similarly y^δ for its perturbation.
Why finite-dimensionality?

$$E[y^\delta - y] = 0 \quad and \quad E[(y^\delta - y)(y^\delta - y)^T] = \sigma^2 I \tag{36}$$

This implies

$$E[\|y^\delta - y\|^2] = m\sigma^2 \quad \Leftrightarrow \quad \delta = \sqrt{m}\sigma$$

The advantage of the white noise assumption is that it allows for a more sophisticated analysis of the propagated data error. With the deterministic approach, so far, we can estimate

$$\left\|x_\alpha^\delta - x_\alpha\right\|^2 = \left\|g_\alpha(K^*K)K^*(y^\delta - y)\right\|^2 \leqslant \left\|\sqrt{\lambda}g_\alpha(\lambda)\right\|^2_{C[0,\|T\|^2]}\delta^2 \tag{37}$$

This upper bound will not always be sharp.Under white noise assumption we get

$$E\left[\left\|x_\alpha^\delta - x_\alpha\right\|^2\right] = \delta^2 trace\{\frac{1}{m}g_\alpha^2(KK^*)KK^*\} \tag{38}$$

The disadvantage - it requires the detailed information about the location of the singular values of K.

In generalized Cross-validation the regualrization parameter α is taken as the minimizer of the functional

$$V(\alpha) = \left(\frac{\|y^\delta - Kx_\alpha^\delta\|}{trace\{\frac{1}{m} r_\alpha(KK^*)\}} \right)^2 \tag{39}$$

in contrast to the previous sections this cross-validation functional is not considered an error estimate, but rather an estimator of the so-called *predictive mean-square error*

$$T(\alpha) = \|y^\delta - Kx_\alpha^\delta\|^2$$

Replacing the regualrization parameter α by the number k of maintained singular values, we get

$$V(k) \approx (1 - \frac{k}{m})^{-2} \left(\|y - Kx_k\|^2 + \delta^2 (1 - \frac{k}{m}) \right) \tag{40}$$

This shows that $V(k)$ attains its minimum for $k << m$.Assuming that $\|y - Kx_k\|^2$ has a resonable decal we get

$$V(k) \approx \|y - Kx_k\|^2 + \delta^2 (1 + \frac{k}{m}) \tag{41}$$

Similarly for $T(k)$ we get,

$$T(k) \approx \|y - Kx_k\|^2 + \delta^2 \frac{k}{m} \tag{42}$$

Comparing these two estimates we conclude that atleast in the range of reasonable regularization parameters $k << m$, $V(k)$ and $T(k)$ behave likewise up to a shift by δ^2.
Consequently , local minima of $T(k)$ in the range are likely to be global minima of $V(k)$.

L-Curve Method

We plot $\|x_\alpha^\delta\|$ versus $\|y^\delta - Tx_\alpha^\delta\|$ in a log-log scale for a large range of α values.

- When α is moderate to large, $\|x_\alpha^\delta\|$ is of the same magnitude of $\|x^\dagger\|$, and it varies comparatively little as α decreases. At the same time $\|y^\delta - Tx_\alpha^\delta\|$ is decreasing with a typical rate of $O(\alpha^\nu)$
- When α is small, the residual norm is of the order of α and changes very little , while the norms of the approximation blow up.

This behavior is shown in the following figure.

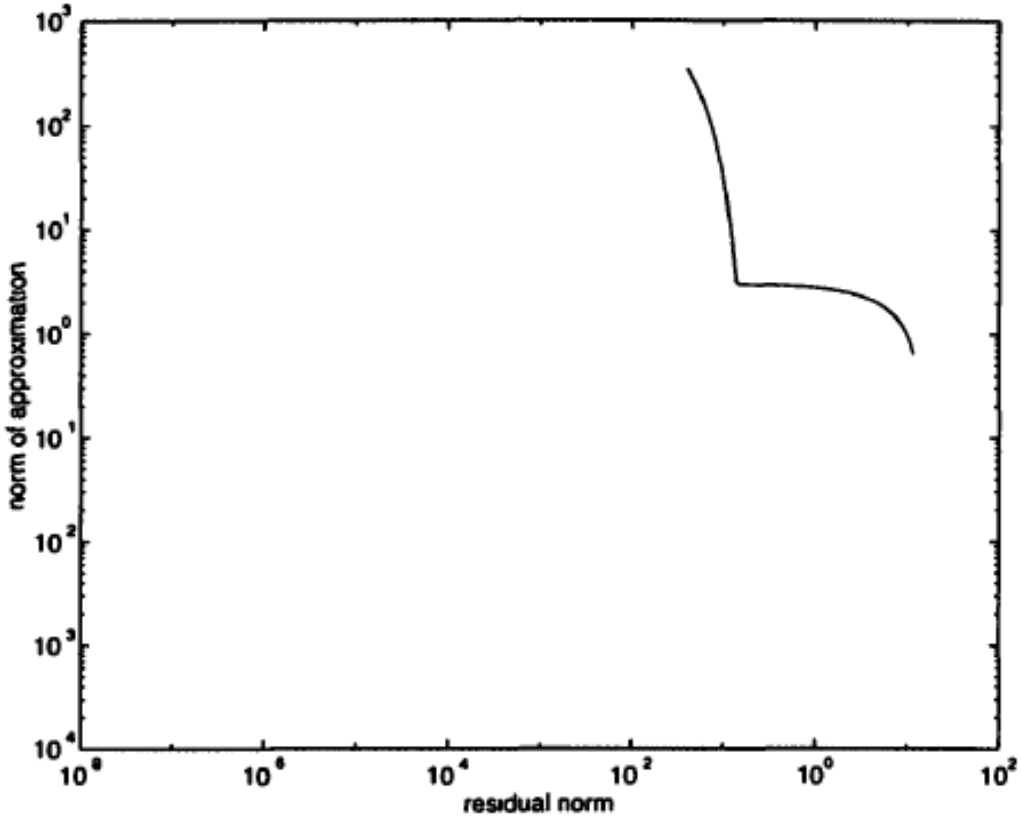

Figure 4.3: L-curve (solid line) made up by "superposition" of two components.

The figure suggest that the horizontal part of the L-curve solely depends on y while the vertical part mainly due to noise.

The crossover point of the two curves is close to the "corner" of the L-curve. Hence the corresponding regularization parameter is a good compromise between data fitting and penalizing the norm of reconstruction.

Hansen and O'Leary Algorithm

The location of the corner should not be choose visually, but rather by some numerical optimization routine.
We seek the point on the graph with maximal curvature.

$$\kappa(\alpha) = \frac{\xi''(\alpha)\eta'(\alpha) - \xi'(\alpha)\eta''(\alpha)}{\left(\xi'(\alpha)^2 + \eta'(\alpha)^2\right)^{3/2}} \tag{43}$$

$$\xi(\alpha) = log\left\|y^\delta - Tx_\alpha^\delta\right\|, \quad \eta(\alpha) = log\left\|x_\alpha^\delta\right\| \tag{44}$$

Reginska's definition of a "corner"

A point $C = (\xi(\alpha_*), \eta(\alpha_*))$ is the corner of the L-Curve if

L is concave in the neighborhood of C, and
the tangent of L at C has the slope -1. (45)

Proposition (4.20)

The point $C = (\xi(\alpha_*), \eta(\alpha_*))$ *is a corner of the L-curve in the aforementioned sense if and only if the function*

$$\psi(\alpha) = \left\|x_\alpha^\delta\right\| \left\|y^\delta - Tx_\alpha^\delta\right\| \tag{46}$$

has a local minimum at $\alpha = \alpha_*$