Statistical Inverse Problems and Instrumental Variables

Thorsten Hohage

Institut für Numerische und Angewandte Mathematik
University of Göttingen

RICAM, Linz, 6.9.2008

outline

nonlinear iterative regularization methods

- preliminaries
- Landweber iteration
- Newton-type iterations

Convergence analysis for known smoothness

Lepskiĭ-type stopping rule

instrumental variables

nonlinear Tikhonov regularization

$$a_\alpha^\delta = \mathrm{argmin}_{a \in D(F)} \left(\|F(a) - Y\|^2 + \alpha \|a - a_0\|^2 \right)$$

consistency and convergence rate analysis:

- for deterministic noise: Engl, Kunisch & Neubauer (1989), Neubauer (1992), Tautenhahn & Qi-nian Jin (2003), ...
- for random noise: O'Sullivan (1990), Bissantz, Hohage & Munk (2004), Loubes & Ludeńa (2005)

difficulties:

- A global minimum is not necessarily unique.
- There may be many local minima.

source conditions for nonlinear problems

Notation:

$F'[a] : \mathcal{X} \to \mathcal{Y}$ Fréchet derivative of F at $a \in D(F)$

$F'[a]^* : \mathcal{Y} \to \mathcal{X}$ adjoint of $F'[a]$

For nonlinear problems source conditions have the form

$$a_0 - a = \Lambda(F'[a]^* F'[a])w \qquad \text{for some } w \in \mathcal{X}$$

where the function Λ is continuous, strictly monotonically increasing, and $\Lambda(0) = 0$.

Again the most important examples are:

Hölder-type source conditions: $\Lambda(t) = t^\nu$, $\nu > 0$

logarithmic source conditions: $\Lambda(t) = (\ln t^{-1})^{-p}$, $p > 0$

nonlinearity vs. degree of ill-posedness

If F' is Lipschitz continuous with Lipschitz constant L (i.e. $\|F'[a_1] - F'[a_2]\| \leq L\|a_1 - a_2\|$ for all a_1, a_2), then the Taylor remainder satisfies the estimate

$$\left\|F(a+h) - F(a) - F'[a]h\right\|_{\mathcal{Y}} \leq \frac{L}{2}\|h\|_{\mathcal{X}}^2, \qquad \|h\| \to 0$$

Consider the linearized equation $F'[a]h = g$, and a regularized solution h_α with regularization parameter α. If the exact solution h satisfies a Hölder source condition with index $\mu > 0$, then the bias $b_\alpha = \mathbf{E}(h_\alpha - h)$ typically fulfills

- $\|F'[a]b_\alpha\|_{\mathcal{Y}} = \mathrm{O}\left(\alpha^{\mu+\frac{1}{2}}\right)$
- $\|b_\alpha\|_{\mathcal{X}}^2 = \mathrm{O}\left(\alpha^{2\mu}\right)$

This estimate of the Taylor remainder is only useful for $\mu \geq \frac{1}{2}$!

iterative regularization methods

An iterative method $a_{k+1}^\delta := \Phi(a_k^\delta, \ldots, a_0, Y)$ together with a stopping rule $K(\delta, \sigma, Y)$ is called an iterative regularization method for estimating a from $Y = F(a) + \delta\zeta + \sigma\xi$ if for all $a \in D(F)$, and all initial guesses a_0 sufficiently close to a the following conditions hold:

- For exact data $(\delta, \sigma = 0)$ either $K = K(\delta, Y) < \infty$ and $a_k^\delta = a$ or $K = \infty$ and $\|a_k - a\| \to 0$ for $n \to \infty$.
- For noisy data the following regularization property is satisfied:

$$\sup_{\|\zeta\| \leq 1} \mathbf{E}\|a_{K(\delta,\sigma,Y)}^\delta - a\|^2 \to 0, \qquad \delta, \sigma \to 0.$$

discrepancy principle

Stop iteration at the first index K for which

$$\|F(a_K^\delta) - Y\|_{\mathcal{Y}} \leq \tau\delta$$

for some fixed $\tau > 1$.

nonlinear Landweber iteration

Idea: Minimize the data misfit functional

$$\Phi(a) := \frac{1}{2}\|F(a) - Y\|^2.$$

Since

$$\Phi'[a]h = \langle F(a) - Y, F'[a]h \rangle = \langle F'[a]^*(F(a) - Y), h \rangle ,$$

the direction of steepest descent is $-F'[a]^*(F(a) - Y)$.
This leads to the iteration formula

$$\widehat{a}_{k+1} = \widehat{a}_k - \mu F'[\widehat{a}_k]^*(F(\widehat{a}_k) - Y)$$

with a step size parameter $\mu > 0$. For Landweber iteration, μ is fixed with $\mu\|F'[a]\|^2 \leq 1$ for all a in a neighborhood of a.

convergence of Landweber iteration

Theorem (Hanke, Neubauer & Scherzer, 1995)

If the nonlinearity condition

$$\|F(a) - F(\bar{a}) - F'[a](a - \bar{a})\|_{\mathcal{Y}} \leq \eta \|F(a) - F(\bar{a})\|_{\mathcal{Y}}$$

holds for all $a, \bar{a}$ in a neighborhood of a and some $\eta < \frac{1}{2}$, then Landweber iteration together with the discrepancy principle with $\tau > 2\frac{1+\eta}{1-2\eta}$ is a regularization method in the sense of the previous definition.

Newton's method

standard Newton method: Compute an update $h_k = a^\delta_{k+1} - a^\delta_k$ by solving the linearized equation

$$F'[\widehat{a}_k]h_k = Y - F(\widehat{a}_k).$$

Since the linearized equation typically inherits the ill-posedness of the nonlinear equation, it must be regularized!

IRGNM and Levenberg-Marquardt

- Iteratively Regularized Gauss-Newton Method (IRGNM): Update $h_k \in \mathcal{X}$ is the unique minimum of the quadratic functional

$$h \mapsto \|F'[\widehat{a}_k]h + F(\widehat{a}_k) - Y\|_{\mathcal{Y}}^2 + \alpha_k \|h + \widehat{a}_k - a_0\|_{\mathcal{X}}^2$$

- Levenberg-Marquardt algorithm: Minimize

$$h \mapsto \|F'[\widehat{a}_k]h + F(\widehat{a}_k) - Y\|_{\mathcal{Y}}^2 + \alpha_k \|h\|_{\mathcal{X}}^2$$

- regularization parameters: $\alpha_k = \alpha_0 q^k$ for some $q \in (0, 1)$

IRGNM for linear problems

- If $F = T$ is linear, $\widehat{a}_{k+1}$ is minimizer of

$$a \mapsto \|Ta - Y\|_{\mathcal{Y}}^2 + \alpha_k \|a - a_0\|_{\mathcal{X}}^2 = \text{min!}$$

- Bias and variance must be balanced by proper choice of the stopping index.
- Optimal rates can only be expected for Hölder smoothness classes with index $\mu \leq 1$!

convergence results for the IRGNM

Theorem (Kaltenbacher, Neubauer, Scherzer '97)

The IRGNM with the discrepancy principle is a regularization method if there exist linear operators $R(\bar{a}, a), Q(\bar{a}, a)$ *and constants* $\gamma_R, \gamma_Q > 0$ *such that*

$$\begin{aligned} F'[\bar{a}] &= R(\bar{a}, a)F'[a] + Q(\bar{a}, a) \\ \|I - R(\bar{a}, a)\| &\leq \gamma_R \|\bar{a} - a\|, \\ \|Q(\bar{a}, a)\| &\leq C_Q \|F'[a](\bar{a} - a)\| \end{aligned}$$

for all all $a, \bar{a}$ *in a ball around a.*

convergence rates:

- Bakushinskii (1992): Hölder source condition with $\mu = 1$
- Kaltenbacher, Neubauer & Scherzer (1997): $0 \leq \mu < 1$:
- Hohage (1997): logarithmic source conditions

IRGNM-type methods with higher qualification

Replace Tikhonov regularization by a linear regularization method with higher qualification:

$$\widehat{a}_{k+1} := a_0 + \Phi_{\alpha_k}(T_k^* T_k)\, T_k^* \Big(Y - F(\widehat{a}_k) + T_k(\widehat{a}_k - a_0) \Big)$$

with $T_k := F'[\widehat{a}_k]$.

- Landweber iteration: Kaltenbacher (1997)
- iterated Tikhonov regularization: Hohage (1999), Kaltenbacher, Neubauer & Scherzer (2005)

outline

nonlinear iterative regularization methods
- preliminaries
- Landweber iteration
- Newton-type iterations

Convergence analysis for known smoothness

Lepskiĭ-type stopping rule

instrumental variables

IRGNM

Iteratively Regularized Gauss-Newton Method (IRGNM): For $\alpha_k = \alpha_0 q^k$ with $q \in (0,1)$ define

$$\widehat{a}_{k+1} := \operatorname{argmin}_{a\in\mathcal{X}} \|F'[\widehat{a}_k](x-\widehat{a}_k)+F(\widehat{a}_k)-Y\|^2_{\mathcal{Y}}+\alpha_k\|a-a_0\|^2_{\mathcal{X}}$$

source conditions:

$$a_0 - a = \Lambda(F'[a]^* F'[a])w \qquad \text{for some } w \in \mathcal{X}.$$

deterministic convergence results:

- Bakushinskii (1992): source condition $\Lambda(t) = t$
- Kaltenbacher, Neubauer & Scherzer (1997): Hölder source conditions $\Lambda(t) = t^\mu$ with $0 \le \mu < 1$
- Hohage (1997): logarithmic source conditions $\Lambda(t) = (-\ln t)^{-p}$, $p > 0$

generalized IRGNM

Consider an arbitrary linear spectral regularization method described by filter functions (g_α) satisfying the usual assumptions.

$$\widehat{a}_{k+1} := a_0 + g_{\alpha_k}\left(F'[\widehat{a}_k]^* F'[\widehat{a}_k]\right) F'[\widehat{a}_k]^* \times \\ \times \left(y^\delta - F(\widehat{a}_k) + F'[\widehat{a}_k](\widehat{a}_k - a_0)\right)$$

For classical IRGNM $g_\alpha(\lambda) = \frac{1}{\lambda+\alpha}$.

Bakushinskii (1995), Kaltenbacher (1997), Hohage (1999), Kaltenbacher, Neubauer & Scherzer (2008)

error decomposition

Let $T := F'[a^\dagger]$ and $T_k := F'[\widehat{a}_k]$. The error $E_k = \widehat{a}_k - a$ in the kth Newton step can be decomposed into

- an approximation error

$$E^{\mathrm{app}}_{k+1} := r_{\alpha_k}(T^*T)E_0,$$

- a propagated data noise error

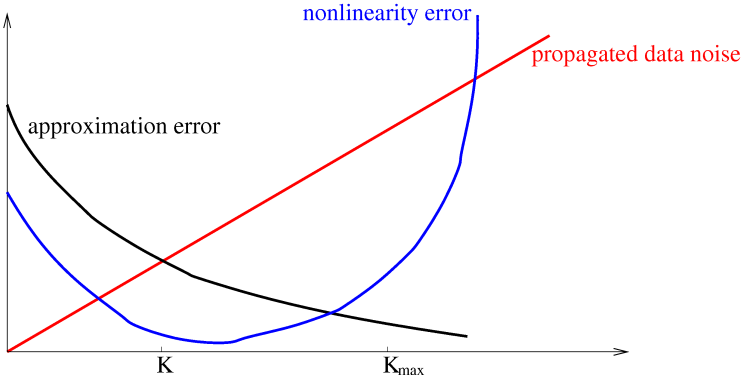
$$E^{\mathrm{noi}}_{k+1} := g_{\alpha_k}(T_k^*T_k)T_k^*(\delta\zeta + \sigma\xi),$$

- and a nonlinearity error

$$\begin{aligned} E^{\mathrm{nl}}_{k+1} &:= g_{\alpha_k}(T_k^*T_k)T_k^*\left(F(a) - F(\widehat{a}_k) + T_kE_k\right) \\ &\quad + \left(r_{\alpha_k}(T_k^*T_k) - r_{\alpha_k}(T^*T)\right)E_0, \end{aligned}$$

i.e. $E_{k+1} = E^{\mathrm{app}}_{k+1} + E^{\mathrm{noi}}_{k+1} + E^{\mathrm{nl}}_{k+1}$.

illustration of the error splitting



aim: show that there exists a sufficiently large, a-priori known $K_{\max} = K_{\max}(\delta, \sigma)$ and $\gamma_{\rm nl} > 0$ such that

$$\|E_k^{\rm nl}\| \le \gamma_{\rm nl} \left(\|E_k^{\rm app}\| + \tau \sqrt{\mathbf{E}\|E_k^{\rm noi}\|^2} \right), \qquad k = 1, \ldots, K_{\max}.$$

approximation and nonlinearity errors

Assumptions:

Hölder source condition: For some $\mu \in \left[\frac{1}{2}, \mu_0\right]$ where μ_0=qualfication we have

$$a_0 - a = (T^* T)^\mu w \qquad \text{with } \|w\| \leq \rho$$

Decay of $\|E_k^{\mathrm{app}}\|$: There exists $\gamma_{\mathrm{app}} > 0$ such that for all k

$$\|E_{k+1}^{\mathrm{app}}\| \leq \|E_k^{\mathrm{app}}\| \leq \gamma_{\mathrm{app}} \|E_{k+1}^{\mathrm{app}}\|$$

This rules out methods with infinite qualification such as Landweber iteration!

Lipschitz condition:
$\|F'[a_1] - F'[a_2]\| \leq L\|a_1 - a_2\|, \qquad a_1, a_2 \in D(F).$

propagated data noise error

We make the following assumptions on the variance term $V(a, \alpha) := \|g_\alpha(F'[a]^* F'[a]) F'[a]^* \xi\|^2$:

- There exists a known function $\varphi_{\rm noi}$ such that
$$(\mathbf{E} V(a, \alpha))^{1/2} \leq \varphi_{\rm noi}(\alpha) \qquad \forall \alpha \in (0, \alpha_0] \text{ and } a \in D(F).$$
- There are constants $1 < \overline{\gamma}_{\rm noi} \leq \overline{\gamma}_{\rm noi} < \infty$ such that
$$\underline{\gamma}_{\rm noi} \leq \varphi_{\rm noi}(\alpha_{k+1})/\varphi_{\rm noi}(\alpha_k) \leq \overline{\gamma}_{\rm noi}, \qquad \forall k \in \mathbb{N}_0.$$
- (exponential inequality)
$\exists \lambda_1, \lambda_2 > 0 \, \forall a \in D(F) \, \forall \alpha \in (0, \alpha_0] \, \forall \tau \geq 1$
$$\mathbb{P}\{V(a, \alpha) \geq \tau \mathbf{E} V(a, \alpha)\} \leq \lambda_1 e^{-\lambda_2 \tau}.$$

Then the "good event" $A_{\tau, \overline{k}}$ that

$$\|E^k_{\rm noi}\| \leq \tau \sigma \varphi_{\rm noi}(\alpha_{k-1}) + \frac{\delta C_g}{\sqrt{\alpha_{k-1}}} =: \Phi_{\rm noi}(k)$$

for $k = 1, \ldots, \overline{k}$ has probabilty $\mathbb{P}(A_{\tau, \overline{k}}) \geq 1 - \overline{k} \lambda_1 e^{-\lambda_2 \tau}$.

crucial lemma

Lemma

Under the assumptions above define

$$K_{\max} := \max\left\{k \in \mathbb{N} : \Phi_{\mathrm{noi}}(k)\alpha_{k-1}^{-1/2} \leq C_{\mathrm{stop}}\right\}$$

$$\text{with } 0 < C_{\mathrm{stop}} \leq \min\left(\frac{1}{8LC_g}, \frac{R}{4\sqrt{\alpha_0}}\right)$$

and assume the "good event" with $\overline{k} = K_{\max}$*. Moreover, let* α_0 *be sufficiently large,* $B_{2R}(a_0) \subset D(F)$*, and define* $\gamma_{\mathrm{nl}} := 8LC_gC_{\mathrm{stop}}$*. Then* $\widehat{a}_k \in B_R(a)$ *and*

$$\|E_k^{\mathrm{nl}}\| \leq \gamma_{\mathrm{nl}}\left(\|E_k^{\mathrm{app}}\| + \Phi_{\mathrm{noi}}(k)\right), \qquad k = 1, \ldots, K_{\max}.$$

This result and the results of the following section are shown in:

F. Bauer, T. Hohage, A. Munk. *Regularized Newton Methods for Nonlinear Inverse Problems with Random Noise.* NAM preprint 2008-06.
ftp://ftp.num.math.uni-goettingen.de/pub/preprints/hohage/main.pdf

convergence for stochastic noise

Theorem

Recall that $E_k^{\mathrm{app}} := r_{\alpha_{k-1}}(F'[a]^* F'[a])(a - a_0)$ *for given* a *and define*

$$K_{\mathrm{or}} := \operatorname{argmin}_{k\in\mathbb{N}} \left(\|E_k^{\mathrm{app}}\| + \sigma \varphi_{\mathrm{noi}}(\alpha_{k-1}) \right).$$

If $\|\widehat{a}_k - a_0\| \leq 2R$ *for* $k = 1, \dots, K$*, set* $K_{\mathrm{or},*} := K$*, otherwise* $K_{\mathrm{or},*} := 0$*. Then*

$$\left(\mathbf{E}\|\widehat{a}_{K_{\mathrm{or},*}} - a\|^2 \right)^{1/2} \leq C \min_{k\in\mathbb{N}} \left(\|E_k^{\mathrm{app}}\| + \sigma \varphi_{\mathrm{noi}}(\alpha_{k-1}) \right).$$

outline

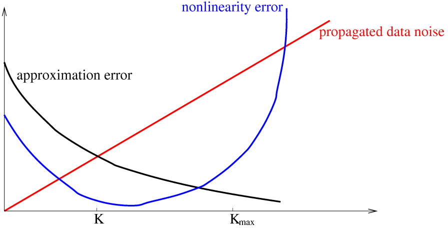
Lepskiĭ balancing principle for nonlinear inverse problems

- Let $\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_{K_{\max}}$ be estimators of a such that

$$\|\widehat{a}_k - a\| \leq \Phi_{\mathrm{noi}}(k) + \Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{nl}}(k), \qquad k \leq K_{\max}.$$

- Φ_{app} is unknown and non-increasing.
- Φ_{noi} is known and non-decreasing.
- Φ_{nl} is unknown and satisfies for some $\gamma_{\mathrm{nl}} > 0$

$$\Phi_{\mathrm{nl}}(k) \leq \gamma_{\mathrm{nl}} \left(\Phi_{\mathrm{noi}}(k) + \Phi_{\mathrm{app}}(k)\right), \quad k = 0, \ldots, K_{\max}.$$



review of standard deterministic theory

source condition: $a = (T^*T)^\mu w$ for some $w \in \mathcal{X}$

smoothness class: $M_{\mu,\rho} := \{(T^*T)^\mu w : \|w\| \le \rho\}$

discrepancy principle: Let $\tau > 1$.

$$\alpha_{\text{disc}} := \sup\{\alpha > 0 : \|Ta_\alpha^\delta - u^\delta\| \le \tau\delta\}$$

Theorem: For $\mu \le \mu_0 - \frac{1}{2}$

$$\|a_{\alpha_{\text{disc}}}^\delta - a\| \le C\rho^{\frac{1}{2\mu+1}}\delta^{\frac{2\mu}{2\mu+1}}$$
$$= C \inf_R \sup\{\|R(Ta+\delta\zeta) - a\| : a \in M_{\mu,\rho}, \|\zeta\| \le 1\}$$

Improved a-posteriori rules with optimal error bounds for $\mu \le \mu_0$ instead of $\mu \le \mu_0 - \frac{1}{2}$:
Raus (1984), Gfrerer (1987), Engl & Gfrerer (1988), Hämarik & Tautenhahn (1999), Mathé & Pereverzev (2003)

oracle inequalities

Let $a \in \mathcal{X}$, $Y = Ta + \sigma\xi$, and $R_\alpha : \mathcal{Y} \to \mathcal{X}$ a family of regularization operators.
An oracle $\alpha_{\rm or}$ is a regularization parameter satisfying

$$\mathbf{E}\|R_{\alpha_{\rm or}}(Y) - a\|^2 = \inf_\alpha \mathbf{E}\|R_{\alpha_*}(Y) - a\|^2$$

An oracle inequality for some given parameter choice rule $\alpha_* = \alpha_*(Y, \sigma)$ has the form

$$\mathbf{E}\|R_{\alpha_*}(Y) - a\|^2 \leq \chi(\sigma)\mathbf{E}\|R_{\alpha_{\rm or}}(Y) - a\|^2,$$

in the optimal case with $\chi(\sigma) = 1 + o(\sigma)$ for $\sigma \to 0$.
Oracle inequalities for statistical inverse problems: Cavalier, Golubev, Picard & Tsybakov (2002); Cavalier & Tsybakov (2002); Cavalier & Golubev (2007)

Deterministic oracle inequality for balancing principle (with constant χ): Mathé & Pereverzev (2006)

comparison of concepts

- An oracle inequality and order optimality for an a-priori rule implies order optimality for an a-posteriori rule.
- The converse is not true, one can even show the following:

Proposition

Let $R_\alpha := (\alpha I + T^*T)^{-1}T^*$ (Tikhonov regularization). Then for every $x \in M_{\mu,\rho}$ we have

$$\sup_{\delta>0} \frac{\rho^{\frac{1}{2\mu+1}}\delta^{\frac{2\mu}{2\mu+1}}}{\inf_\alpha \sup\{\|R_\alpha(Ta+\delta\zeta)-a\| : \|\zeta\|\leq 1\}} = \infty$$

In other words: For every element x in the smoothness class $M_{\mu,\rho}$ there exists an error level $\delta > 0$ for which the classical deterministic error bounds are suboptimal by an arbitrarily large factor!

oracle inequality

Lepskiĭ balancing principle:

$$k_{\mathrm{bal}} := \min\left\{k \le K_{\max} : \begin{array}{l} \|\widehat{a}_k - \widehat{a}_m\| \le 4(1+\gamma_{\mathrm{nl}})\Phi_{\mathrm{noi}}(m), \\ m = k+1, \dots, K_{\max} \end{array}\right\}$$

Theorem

Assume that $\Phi_{\mathrm{noi}}(k+1) \le \overline{\gamma}_{\mathrm{noi}}\Phi_{\mathrm{noi}}(k)$ *for some constant* $\overline{\gamma}_{\mathrm{noi}} < \infty$. *Then*

$$\|\widehat{a}_{k_{\mathrm{bal}}} - a\| \le 6(1+\gamma_{\mathrm{nl}})\overline{\gamma}_{\mathrm{noi}} \min_{k=1,\dots,K_{\max}} \left(\Phi_{\mathrm{app}}(k) + \Phi_{\mathrm{noi}}(k)\right).$$

special case $\gamma_{\mathrm{nl}} = 0$: Mathé & Pereverzev (2006); Mathé (2006)

see also Lepskiĭ (1990); Goldenshluger & Pereverzev & Schock (2000); Mathé & Pereverzev (2003); Pereverzev & Schock (2005)

deterministic errors, unknown smoothness

Corollary

Let $u = F(a) + \delta\zeta$. *Then*

$$\|\widehat{a}_{k_{\mathrm{bal}}} - a\| \quad \leq \quad 6(1+\gamma_{\mathrm{nl}})\overline{\gamma}_{\mathrm{noi}} \inf_{k\in\mathbb{N}} \left(\Phi_{\mathrm{app}}(k) + \frac{\delta C_g}{\sqrt{\alpha_{k-1}}} \right)$$

In particular, for a Hölder source condition with index $\mu \geq 1/2$ *we have*

$$\|\widehat{a}_{k_{\mathrm{bal}}} - a\| = \mathrm{O}\left(\rho^{\frac{1}{2\mu+1}} \delta^{\frac{2\mu}{2\mu+1}} \right)$$

stochastic noise, unknown smoothness

Corollary

Let $\delta = 0$ and $\tau = \ln \sigma^{-2}$, i.e.

$$\Phi_{\mathrm{noi}}(k) := (\ln \sigma^{-2})\sigma\varphi_{\mathrm{noi}}(\alpha_{k-1}).$$

Furthermore let k_{bal} be chosen by the Lepskiĭ balancing principle if $\widehat{a}_k \in B_{2R}(a_0)$ for $k = 1, \dots, K_{\max}$ and $k_{\mathrm{bal}} := 0$ else. Then there exists a constant $C > 0$ such that for σ small enough and $\mu > \frac{1}{2}$

$$\left(\mathbf{E}\|\widehat{a}_{k_{\mathrm{bal}}} - a\|^2\right)^{1/2} \leq C \min_{k\in\mathbb{N}} \left(\Phi_{\mathrm{app}}(k) + (\ln \sigma^{-2})\sigma\varphi_{\mathrm{noi}}(\alpha_{k-1})\right).$$

outline

nonlinear iterative regularization methods
- preliminaries
- Landweber iteration
- Newton-type iterations

Convergence analysis for known smoothness

Lepskiĭ-type stopping rule

instrumental variables

introduction

- regression problem: Estimate a function a given n independent observations (X_i, Z_i), $i = 1, \dots, n$ of random variables X, Z satisfying

$$Z = a(X) + \epsilon$$

  where ϵ is an unobservable nuisance variable satisfying $\mathbf{E}(\epsilon|X) = 0$.
- Often the assumption $\mathbf{E}(\epsilon|X) = 0$ is violated.
- We will show that by solving an ill-posed inverse problem one can still estimate a if there exists another *observable* quantity W, which is sufficiently correlated with X and satisfies

$$\mathbf{E}(\epsilon|W) = 0.$$

Estimating hourly wages as a function of the education level

- Z_i: hourly wage of indiviual i
- X_i: level of education of individual i
- unknown: $a(X) := \mathbf{E}(Z|X)$

Here it seems unlikely that the wage X and the nuisance variable $\epsilon = Z - a(X)$ are uncorrelated since there are other variables such intellegence and stamina which influence both X and Z.
However, we may choose W e.g. as distance of the individuals appartment from college and reasonably assume that $\mathbf{E}(\epsilon|W) = 0$.

a linear first kind integral equation

From the observed data (X_i, Z_i, W_i) we can estimate the joint density $f(x, y, w)$. We have

$$\int_X \mathbf{E}(X = x|W = w)g(x)dx = \mathbf{E}(Z|W = w) \qquad \text{for all } w.$$

Setting $k(w, x) := \mathbf{E}(X = x|W = w)$ and $u(w) := \mathbf{E}(Z|W = w)$ we obtain the linear integral equation

$$\int k(w, x)g(x)\, dx = u(w).$$

Note that both the right hand side and the kernel are noisy since they have to be estimated from the data.

P. Hall and J.L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Ann. Stat., 33: 2904–2929*, 2005.

a nonlinear integral equation

Often the assumption $\mathbf{E}(X|W) = 0$ can be replaced by the stronger independence assumption

$$\epsilon, W \text{ independent}, \qquad \mathbf{E}\epsilon = 0.$$

The first assumption is equivalent to

$$\int f(\epsilon + a(x), x, w)\, dx = \int f_W(\epsilon + a(x), x) f_{Y,X}(w)\, dx \quad \text{for all } \epsilon, w$$

where f_W and $F_{Y,X}$ denote the marginal densities w.r.t. W and Y, X, respectively.
This is a nonlinear integral equation with a noisy kernel, which can be solved by regularized Newton methods.

joint work with J.P.Florens, J. Johannes and E. Mammen