Regularization independent of the noise level: an analysis of quasi-optimality

Frank Bauer¹ and Markus Reiß²

¹ Fuzzy Logic Laboratorium Linz-Hagenberg, Johannes Kepler University of Linz, Softwarepark 21, 4232 Hagenberg, Austria
² Institute for Applied Mathematics, University of Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany

E-mail: frank.bauer@jku.at and reiss@statlab.uni-heidelberg.de

Received 21 June 2007, in final form 27 July 2008
Published 19 August 2008
Online at stacks.iop.org/IP/24/055009

Abstract
The quasi-optimality criterion chooses the regularization parameter in inverse problems without taking into account the noise level. This rule works remarkably well in practice, although Bakushinskii has shown that there are always counterexamples with very poor performance. We propose an average case analysis of quasi-optimality for spectral cut-off estimators (also known as truncated singular value decomposition, TSVD) and we prove that the quasi-optimality criterion determines estimators which are rate-optimal on average. Its practical performance is illustrated with a calibration problem from mathematical finance.

1. Introducing
We consider the prototype of a linear inverse problem where we observe \( y = Ax + \xi \) with a compact operator \( A \) on a Hilbert space \( X \) and some noise variable \( \xi \) and where we try to recover a stable approximation of the true solution \( x \in X \). In this setting we address the question, why the so-called quasi-optimality criterion for choosing the regularization parameter in stable inversion algorithms, as proposed by Tikhonov and Arsenin (1977) and Tikhonov et al (1979), works remarkably well in many practical situations.

The classical quasi-optimality criterion is applied to the regularized solutions
\[
x_\alpha = (A^*A + \alpha I)^{-1}A^*y, \quad \alpha > 0,
\]
based on Tikhonov’s method and proposes to choose \( \alpha > 0 \) such that
\[
\left\| \frac{dx_\alpha}{d\alpha} \right\| \rightarrow \min! \quad \alpha
\]
Reparametrizing with $\alpha = q^n$ for $q \in (0, 1)$ and $n \in \mathbb{R}$ yields $\| \frac{dx_n}{dn} \| \to \min_n$. When for practical purposes only a discrete grid $\{q^n | n \in \mathbb{N}\}$ is considered, then this reduces to the following criterion:

$$\| x_{q^n} - x_{q^{n+1}} \| \to \min_n$$

Obviously this parameter choice rule does not rely on any knowledge concerning the operator, the solution or the noise.

On the other hand, Bakushinskii (1984) has shown for deterministic noise $\xi$ that a method for choosing the regularization parameter should depend on the noise level $\delta \geq \| \xi \|$, whereas quasi-optimality does not.

**Theorem 1.1** (Bakushinskii 1984). If the regularized solution operator $R : X \to X$ does not depend explicitly on the noise level $\delta$, then for any (infinite rank) compact operator $A : X \to X$ there is a $y \in \text{Dom}(A^+) = \text{Dom}(A^{-\frac{1}{2}})$ for which

$$\lim_{\delta \to 0} \sup_{y \in X} \frac{\| R_y - A^+y \|}{\| y - y_\delta \|} > 0,$$

where $A^+$ denotes the generalized inverse of $A$. Bakushinskii’s result is based on a worst-case scenario or minimax paradigm: for any such method $R$ there exist solution pairs $(x, y)$, where the method does not converge uniformly over all perturbed observations $y_\delta$. These important counterexamples are artificially constructed and need not at all be typical. The core message of our paper is in fact that for the quasi-optimality method these counterexamples are usually rare and we can even get optimal or near to optimal solutions without knowing the noise level $\delta$ when applying an average case analysis.

Specifying an average case scenario amounts to prescribing a weighting scheme or— in statistical terms—a Bayesian a priori law for the true solutions $x$. We suppose that the coefficients of the solution in the singular value decomposition are normally distributed around zero. Only a very general condition on the decay property of the variances will be imposed. Then we can prove our main theorem 3.4 that estimators, based on a truncated singular value decomposition and the quasi-optimality criterion for selecting the cut-off value, are optimal on average. Our analysis also applies to variants of quasi-optimality like hardened balancing (Bauer 2007), which are sometimes preferable in practice.

First considerations, indicating why this kind of methods might work, are already given by Bauer (2007). We refer to the monograph Engl et al (1996) for a general overview about practical inverse problems, their abstract mathematical formulation and methods for choosing the regularization parameter. The Bayesian framework, which is related to our approach, is discussed by Kaipio and Somersalo (2005). The choice of the regularization parameter is, of course, a perennial problem in nonparametric statistics, see, e.g., Wasserman (2006), and diverse methods have been applied to statistical inverse problems, e.g. generalized cross validation (Wahba 1977), Lepski’s method (Lepski 1990, Bauer and Pereverzev 2005) or wavelet thresholding (Cohen et al 2004), which depend, however, heavily on the knowledge of the noise level. In this stochastic formulation the noise level $\delta$ is defined slightly differently. Under a white noise assumption, the Fourier coefficients $\langle \xi, u_k \rangle$ with respect to an orthonormal basis $\{u_k | k \in \mathbb{N}\}$ are normal, independent and identically distributed (i.i.d.) with mean zero and variance $\delta^2$. The user-friendly and heuristically motivated quasi-optimality criterion and its variants are attractive in practice because they do not rely on the knowledge of this noise level.

The precise setting is described in section 2, which is followed by the mathematical analysis in section 3. The proofs are postponed to section 6. The numerical example in
section 4 treats the calibration of an option price model from mathematical finance. In this typical inverse problem in financial engineering the choice of the regularization parameter is extremely difficult and the proposed methods yield comparatively good results. A short outlook in section 5 concludes this paper.

2. The framework

2.1. Observations and estimators

We consider $A : X \rightarrow X$, a compact, self-adjoint and (strictly) positive-definite operator on the real Hilbert space $X$ with singular value decomposition

$$Ax = \sum_{k=1}^{\infty} \lambda(k)x_k u_k = \sum_{k=1}^{\infty} \lambda(k)x_k u_k,$$

where $(u_k)$ is an orthonormal basis of eigenvectors and the positive eigenvalues $(\lambda(k))_k$ are arranged in decreasing order, satisfying $\lim_{k \to \infty} \lambda(k) = 0$. From the general statistical observation model

$$y = Ax + \xi, \quad \xi \text{ noise},$$

we immediately go over to a sequence space model by considering the coordinates with respect to $(u_k)$. We observe

$$y_k = \lambda(k)x_k + \varepsilon \xi_k, \quad k \geq 1.$$

Here, $\xi_k$ are assumed to be independent and identically distributed (i.i.d.) standard normal random variables and $(x_k)$ the coordinates of the unknown quantity $x$, which is to be estimated. Writing $\sigma(k) := \varepsilon(k)/\lambda(k)$, the empirical coefficients of $x$ are given by

$$\tilde{x}_k := \lambda(k)^{-1}y_k = x_k + \varepsilon k^\frac{1}{2}\lambda^{-1}x_k = x_k + \sigma(k)x_k, \quad k \geq 1. \quad (2)$$

Equation (2) is the abstract observation model we shall consider from now on. Let us point out that we shall not speak of a noise level in the observations since the noise intensity is described by the function $\varepsilon(\cdot)$ which is integrated in the function $\sigma(\cdot)$. The function $\sigma(\cdot)$ is now referred to as noise behavior, which directly indicates the ill-posedness of the inverse problem due to $\sigma(k) \rightarrow \infty$ for $k \rightarrow \infty$. For $\varepsilon(\cdot) = \delta$ we have exactly the definition of white noise as discussed above.

We shall use a subsampling function $\ell : \mathbb{N} \rightarrow \mathbb{N}$ with $\ell(n + 1) > \ell(n)$. Applying a spectral cut-off scheme, our estimator of $x$ at level $n$ for a given subsampling $\ell$ is defined by the truncated singular value decomposition

$$\hat{x}^{(n)} := \sum_{k=1}^{\ell(n)} \tilde{x}_k u_k.$$

Its mean-squared error (MSE) is given by the bias–variance decomposition, see, e.g., Wasserman (2006),

$$\mathbb{E}[\|\hat{x}^{(n)} - x\|^2] = \sum_{k=1}^{\ell(n)} \sigma(k)^2 + \sum_{k=\ell(n)+1}^{\infty} \lambda_k^2.$$ 

Recall that the bias–variance dilemma is the fact that the variance, i.e. the first sum, increases with $n$, while the squared bias, i.e. the second sum, decreases. The value of $n$ where the total sum is minimal depends on bias and variance and thus on the properties of the unknown $x$ and on the noise amplification $\sigma(\cdot)$. The quasi-optimality criterion gives a data-driven choice for $n$. 3
2.2. Assumptions

Let us now adopt a Bayesian point of view and perform an average case analysis. We weight the coefficients of \( x \) by the prior distribution

\[
x_k \overset{\text{i.i.d.}}{\sim} N(0, \gamma(k)^2), \quad k \geq 1.
\]

Assumption 2.1 will implicitly require certain decay properties of the variances \( \gamma(k)^2 \) for \( k \to \infty \), but we need not know them in detail. Writing \( \tilde{E} \) for the joint expectation with respect to \( (\xi(k)) \) and \( (x_k) \), the Bayesian risk \( R_2 \) for the MSE is given by

\[
R_2(\hat{x}(n)) = \tilde{E}[\|\hat{x}(n) - x\|^2] = \sum_{k=1}^{\ell(n)} \sigma(k)^2 + \sum_{k=\ell(n)+1}^{\infty} \gamma(k)^2.
\]

\[
s(n) := \sum_{k=1}^{\ell(n)} \sigma(k)^2 \quad \text{(variance),}
\]

\[
b(n) := \sum_{k=\ell(n)+1}^{\infty} \gamma(k)^2 \quad \text{(mean squared bias)}
\]

yields the risk decomposition

\[
R_2(\hat{x}(n))^2 = s(n) + b(n).
\]

**Assumption 2.1** (geometric growth, decay rates). We assume that with some \( 1 < c_s \leq C_s \) and \( 1 < c_b \leq C_b \) we have for all \( n \geq 1 \)

\[
c_s s(n) < s(n + 1) < C_s s(n), \quad c_b b(n + 1) < b(n) < C_b b(n + 1).
\]

The strict inequality in the assumption implies that \( b(n) > 0 \) holds for all \( n \geq 1 \) such that any specific solution \( x \) like \( x = 0 \) has zero prior probability.

Another important observation is that multiplicative changes \( \omega > 0 \) of the noise, i.e. \( \epsilon(\cdot) \to \omega\epsilon(\cdot) \) do not change the constants above. Hence in the white noise case we can say that they are independent of the actual noise level.

**Example 2.2.** This assumption can be easily satisfied for most moderately ill-posed problems with Hölder source conditions (Engl et al 1996) by choosing an exponential subsampling function \( \ell \). Assume that we have \( \lambda(k) \sim k^{-\nu}, \nu > 0 \), and \( \gamma(k) \sim k^{-\mu}, \mu > 1/2 \) and white noise of level \( \delta > 0 \) (i.e. \( \epsilon(k) = \delta \)), and hence \( \sigma(k) = \delta \lambda(k)^{-1} \sim \delta k^\nu \). The white noise setting is assumed for simplicity here, but is not strictly necessary although it gives us access to a canonical notion of a noise level. An exponential subsampling like \( \ell(n) \in [\ell(0)h^n] \) with \( \ell(0) \in \mathbb{N}, h > 1 \) independent of the noise level \( \delta \) and the source conditions, yields

\[
s(n) = \sum_{k=1}^{\ell(n)} \sigma(k)^2 \sim \sum_{k=1}^{\ell(0)h^n} \delta^2 k^{2\nu} \sim \frac{\delta \ell(0)h^{2\nu+1}}{2^\nu+1} (h^{2\nu+1})^n
\]

\[
b(n) = \sum_{k=\ell(n)+1}^{\infty} \gamma(k)^2 \sim \sum_{k=\ell(0)h^n+1}^{\infty} k^{-2\mu} \sim \frac{\ell(0)h^{2\mu+1}}{2^{\nu+1}} (h^{2\mu+1})^n,
\]

and hence \( s(n) \) and \( b(n) \) fulfill assumption 2.1.

Let us further introduce a weight function \( \chi \), which will allow us to treat also variants of quasi-optimality, cf remark 2.6.
Assumption 2.3 (weight function). Let $\chi : \mathbb{N} \to \mathbb{R}$ be a weight function which satisfies for some constants $c_{\chi}, C_{\chi}$ with $c_{\chi}^{-1} < c_{\chi} \leq C_{\chi} < c_{\chi}$,

$$c_{\chi}\chi(n+1) \leq \chi(n) \leq C_{\chi}\chi(n+1).$$

Example 2.4. For $\chi(\cdot) = 1$ assumption 2.3 is obviously fulfilled because of assumption 2.1 using $c_{\chi}^{-1} < 1 = c_{\chi} = C_{\chi} = 1 < c_{\chi}$. Using that $s(n)$ is increasing, assumption 2.3 also holds for $\chi(n) = s(n)^{-1/2}$ using assumption 2.1 with $c_{\chi}^{-1} < 1 = c_{\chi} \leq C_{\chi} = C_{s}^{1/2} < c_{s}$ as long as the constants $C_{\chi}$ and $c_{s}$ are close enough together fulfilling $C_{s}^{1/2} < c_{s}$.

2.3. Choosing the regularization parameter

Definition 2.5. Given a weight function $\chi$, we choose the cut-off level in a data-driven way according to the minimum distance or quasi-optimality criterion (Tikhonov and Arsenin 1977, Tikhonov et al. 1979),

$$n^* := \arg\min_{n \geq 1} \{\chi(n) \|\hat{x}(n+1) - \hat{x}(n)\|_2\}. \quad (3)$$

Remark 2.6. For $\chi(\cdot) = 1$ this is a discretized version of the quasi-optimality criterion; for $\chi(n) = s(n)^{-1/2}$ a version of the hardened balancing principle (Bauer 2007). In theorem 3.4, it is proved that the infimum of the criterion over $n$ is almost surely attained and thus $n^*$ is well defined. It is unique when we take as minimizer the smallest index $n$ where the minimum is attained.

Obviously, we do not need at any point an explicit or implicit knowledge of the noise behavior for computing $\|\hat{x}(n+1) - \hat{x}(n)\|_2$.

The question remains how to minimize the criterion numerically. First of all, we have in practice an idea about a lower bound for the noise. Furthermore, in practical applications we only have a finite number of observations and we shall never deal with more coefficients in the sequence space model than observations available.

3. The mathematical analysis

In this section, we will formalize the following heuristic observation (here case $\chi(\cdot) = 1$). By the geometric decay and growth in assumption 2.1,

$$\mathbb{E}[\|\hat{x}^{(n)} - \hat{x}^{(n+1)}\|^2] = (s(n+1) - s(n)) + (b(n) - b(n+1)) \asymp s(n) + b(n) = R2(\hat{x}^{(n)})^2.$$

Hence, minimizing the criterion is related to minimizing the average case risk. Note, however, that a careful analysis is needed since the criterion is random and not at all independent of the estimator.

In this section, we first determine a rate-optimal estimator for different moment-type losses, based on the (unrealistic) knowledge of the index $n^*$ where the curves $b(\cdot)$ and $s(\cdot)$ intersect. Then we show that this rate does not deteriorate when we take the estimator with the data-driven index $n^*$, based on the quasi-optimality criterion of definition 2.5. All proofs are deferred to section 6.
3.1. Risk for different moments

We denote the normalized Bayes risk for the moments of order $\alpha > 0$ by
\[
R_\alpha(\hat{x}(n)) := \tilde{E}[\|\hat{x}(n) - x\|^\alpha]^{1/\alpha},
\]
consistent with the definition of $R_2(\hat{x}(n))$. Note the order $R_\alpha(\hat{x}(n)) \leq R_\beta(\hat{x}(n))$ for $\alpha \leq \beta$.

**Definition 3.1.** Let $n^*$ be the index where $s(n^* + 1) > b(n^* + 1)$ and $s(n^*) \leq b(n^*)$.

Note that $n^*$ is uniquely defined by the monotonicity properties of $s(\cdot)$ and $b(\cdot)$. Interestingly, for $n^*$ the risk is up to a multiplicative factor minimal, even under different moments:

**Lemma 3.2.** Denote by $\Gamma_1$ the gamma function and by $\zeta$ a standard normal random variable.

With $K_\alpha := \sqrt{2/c}(4\Gamma\left(\frac{\alpha}{2} + 1\right))^{1/\alpha} e^{-E[\log(\|\zeta\|)]}$ the following bounds hold for all $\alpha > 0$:
\[
\forall n \in \mathbb{N} : R_\alpha(\hat{x}(n)) \geq e^{-E[\log(\|\zeta\|)]} R_2(\hat{x}(n)), \\
\forall n \in \mathbb{N} : R_\alpha(\hat{x}(n)) \leq (4\Gamma\left(\frac{\alpha}{2} + 1\right))^{1/\alpha} R_2(\hat{x}(n)), \\
\min_n R_\alpha(\hat{x}(n)) \leq R_\alpha(\hat{x}(n^*)) \leq K_\alpha \min_n R_\alpha(\hat{x}(n)).
\]

We conclude from the last property that the risk (the expectation of the error) at the intersection point $n^*$ of $s(\cdot)$ and $b(\cdot)$ is bounded by a constant multiple of the minimal possible risk. The quasi-optimality criterion aims at approximating this index $n^*$ as stably as possible.

3.2. Main result

First, we quantify the probability $\tilde{P}(n^* = n)$ for any $n \in \mathbb{N}$.

**Proposition 3.3.** Grant assumptions 2.1 and 2.3. Then
\[
\tilde{P}(n^* = n) \leq (\sqrt{2} + (2er)^{1/2})(\rho(n) \log(\rho(n)^{-1}))^{r/2},
\]
where
\[
r := \inf_n r(n) \quad \text{with} \quad r(n) := \frac{\|(s(n + 1) - s(n)) + (b(n) - b(n + 1))\|}{\max_{k(n) < k \leq (n + 1)}(\sigma(k)^2 + \gamma(k)^2)} \geq 1
\]
and
\[
\rho(n) := \frac{c_s c_{\chi}^{-1} (c_k C_{\chi}^{-1})^{n - n^*}}{c_s c_{\chi}^{-1} (c_k C_{\chi}^{-1})^{n - n^*}}, \quad \text{for} \quad n \geq n^*, \\
\rho(n) := \frac{c_s c_{\chi}^{-1} (c_k C_{\chi}^{-1})^{n - n^*}}{c_s c_{\chi}^{-1} (c_k C_{\chi}^{-1})^{n - n^*}}, \quad \text{for} \quad n < n^*.
\]

Since $\rho(n)$ decays exponentially fast, this result yields a rapid decay for the probability that the minimal value for the quasi-optimality criterion is obtained away from the index $n^*$. The slope of the exponential is largely dependent on the subsampling function used, which will be investigated further in the upcoming example 3.6. We are now prepared to state and prove our main theorem.

**Theorem 3.4.** Grant assumptions 2.1 and 2.3. Assume that $\alpha > 0$ satisfies
\[
\alpha < \max_r \left(\frac{\log(C_{\chi})}{\log(c_s C_{\chi}^{-1})}, \frac{\log(C_{\chi})}{\log(c_s c_{\chi}^{-1})}\right).
\]
Then our estimator is almost surely well defined, i.e. the minimum in definition 2.5 for \( n^\ast \) is indeed attained, and it satisfies the oracle-type inequality

\[
R_n(\hat{x}(n^\ast)) \leq K \min_n R_n(\hat{x}(n))
\]

with a constant \( K = K(c_s, c_b, C_s, C_b, c_X, C_X, r, \alpha) > 0 \).

In any case, we have \( r \geq 1 \) by definition, and hence there is a moment \( \alpha > 0 \) for which the quasi-optimality criterion always works.

**Remark 3.5.** A closer look at the proof of proposition 3.3 reveals that for this result only the boundedness and the decay behavior of the Gaussian density was used. Hence, a similar result will hold for more general noise densities with exponential decay.

This oracle-type result, see, e.g., Wasserman (2006) for a general discussion, is actually better than the classical statement obtained in inverse problems which is mostly a worst case estimate. There the results are given in the form (Engl et al. 1996)

\[
\|x - \hat{x}(n^\ast)\| \leq c\delta^\tau,
\]

where \( \tau \) depends on the type of noise and the source conditions. When \( x \) or the noise has ‘accidentally’ better properties than expected we still have this bound although \( \min_n \|x - \hat{x}(n)\| \ll c\delta^\tau \) for \( \delta < \delta_0 \).

In our case of an oracle inequality this cannot happen because we always compare with the best possible (‘oracle’) solution. So we still have the theoretical upper bound in the form \( c\delta^\tau \), but can additionally claim that (on average) we are very close to the best possible solution, even if it is much better than any specific rate derived.

Although the mathematical result depends on a number of constants, it is important to note that the existence of these constants is required to derive the mathematical result, their actual values are not at all used for the quasi-optimality criterion.

In particular for a white noise model of level \( \delta \), these constants do not depend on the value \( \delta \). Our estimator and the oracle inequality (4) are thus completely independent of the noise behavior. This noise behavior enters only implicitly into the error estimate via the oracle risk \( \min_n R_n(\hat{x}(n)) \).

It is interesting to note that though we face a stochastic noise we do not lose a logarithmic factor in the data-driven method as is sometimes the case for other parameter choice methods (Cohen et al. 2004, Bauer and Pereverzev 2005).

The general picture is that our data-driven choice \( n^\ast \) of the cut-off value yields an optimal risk bound up to constants whenever the moments taken are not so large. In order to achieve higher moments, we need to choose a subsampling function \( \ell(\cdot) \) that makes \( r \) large enough.

In practice, we fix the subsampling \( \ell(\cdot) \) and can thus be sure to have an error bound for some moment \( \alpha > 0 \), but as the next example shows, we can also fix the moment \( \alpha \) we want to attain and then determine a good subsampling.

**Example 3.6.** For the standard case of quadratic risk we need \( \alpha = 2 \). For very tight exponential bounds \( C_s \approx c_s, C_b \approx c_b \), it suffices to have \( r > 2 \). Let us reconsider example 2.2, where

\[
(s(n + 1) - s(n)) + (b(n) - b(n + 1))
\approx \frac{\delta^2 \ell(0)^{2\nu+1}}{2\nu + 1} (h^{2\nu+1})^\nu (h^{2\nu+1} - 1) + \frac{\ell(0)^{-2\mu-1}}{2\mu - 1} (h^{-2\mu+1})^{\nu+1} (h^{2\mu-1} - 1).
\]
Hence we have

\[
    r(n) \approx \min \left\{ \frac{\ell(0)(\ell^{(0)})^2}{2^{n+1}(h^{2n+1} - 1)} \left( \frac{\ell(0)(\ell^{(0)})^2}{2^{n+1}(h^{2n+1} - 1)} \right)^{2\nu}, \frac{\ell(0)(\ell^{(0)})^2}{2^{n+1}(h^{2n+1} - 1)} \right\}
\]

\[
    = \ell(0) \min \left\{ \frac{1}{2^{\nu+1}(h^{2n+1} - 1)}, \frac{1}{2^{\mu+1}(h^{2n+1} - 1)} \right\}
\]

\[
    \geq r(1) \approx \ell(0) \min \left\{ \frac{1}{2^{\nu+1}(h^{2n+1} - 1)}, \frac{1}{2^{\mu+1}(h^{2n+1} - 1)} \right\}
\]

\[
    \approx \ell(0)h(h-1) \min \left\{ \frac{1}{2^{\nu+1}}, \frac{1}{2^{\mu+1}} \right\} = (\ell(2) - \ell(1)) \min \left\{ \frac{1}{2^{\nu+1}}, \frac{1}{2^{\mu+1}} \right\}.
\]

Thus the first (and smallest) distance \( \ell(2) - \ell(1) \) needs to be at least 3 in this case. This finding corresponds very well with numerical experience, where it seems that we do not run into major problems even for this relatively small step size.

It is, however, important to note that there is a dependence on the smoothness of the solution (the operator can normally assumed to be known); if we are interested in a quadratic risk and have an initial spacing of \( \ell(2) - \ell(1) \) we attain optimal rates only for all solutions \( x \) where the exponent in the decay rate of the coefficients obeys \( \mu \leq \mu_0 \) with some \( \mu_0 = \mu_0(C_1, C_2, \ldots, C_n, \alpha, \ell(2) - \ell(1)) \). This behavior is reminiscent of the so-called qualification in other regularization methods.

4. An application

In Bauer (2007) the quasi-optimality, Lepski balancing and hardened balancing principles have been compared numerically. The findings are that both, for a large scale stochastic experiment inverting ill-conditioned matrices and for a more realistic experiment determining the field of gravity from satellite data, quasi-optimality and hardened balancing perform quite well and stably, in particular better than the Lepski balancing principle.

The advantage of quasi-optimality is that it can cope with a largely unknown structure of the noise. Therefore, we will present in the sequel numerical experiments for an inverse problem arising in option pricing where noise enters from various sources. Its level is not easy to estimate, furthermore we do not have a good description of the underlying noise model.

The calibration of financial models based on option prices has attracted increasing attention recently due to its practical importance and mathematical challenges, see, e.g., Crépey (2003) and Egger et al (2006) and the references therein for the case of a generalized Black–Scholes model and chapter 13 in Cont and Tankov (2004) for jump models as considered here.

4.1. The calibration problem

We consider the problem of calibrating an exponential Lévy model based on market prices of European options and closely follow Belomestny and Reiß (2006). We briefly describe the problem, but refer to Cont and Tankov (2004) for a thorough introduction. It is assumed that we observe prices \( C(K_j) \) of European call options with different strike prices \( K_j, j = 1, \ldots, n \), and same time \( T > 0 \) to maturity and that the underlying stock price follows an exponential Lévy process

\[
    S_t = S_0 e^{L_t} \text{ with a Lévy process } (L_t).
\]
Here, the Lévy process is restricted to be a superposition of a Brownian motion of volatility \( \sigma^2 > 0 \), a linear drift of slope \( \gamma \in \mathbb{R} \) and a compound Poisson jump process of intensity \( \lambda > 0 \) with jump density \( \nu : \mathbb{R} \to \mathbb{R}^+ \). The goal is to estimate these model parameters, in particular the jump density, which because of only finitely many observations and the presence of bid-ask spreads is a typical inverse problem with noisy observations in quantitative finance. The knowledge of these parameters permits to get a clear picture of the expectations at the market concerning future jumps in the stock price, which is essential for well-founded risk management and pricing of path-dependent options.

We transform the strike prices \( (K_j) \) to the so called log-forward moneyness \( (\chi_j) \) and the call option prices \( \mathcal{O}(K_j) \) to a better behaved generalized option price function \( \mathcal{O}(\chi_j) \) and we introduce the weighted jump density \( \mu(x) = e^{\gamma(x)} \). Then the forward formula, expressing option prices in terms of the model parameters, is given in the spectral domain by \( (\mathcal{F} \text{ denotes the Fourier transform})
\[
\mathcal{F} \mathcal{O}(\nu) = \frac{1 - \exp(T(-\sigma^2(v-i)^2/2 + i\gamma(v-i) + \mathcal{F} \mu(v) - \lambda))}{v(v-i)}, \quad v \in \mathbb{R},
\]
\]

cf equation (2.7) in Belomestny and Reiß (2006). Our observations are modeled as
\[
\mathcal{O}_j = \mathcal{O}(\chi_j) + \varepsilon_j, \quad j = 1, \ldots, n,
\]

with i.i.d. and centered noise variables \( (\varepsilon_j) \). For the sake of an easier presentation here, we assume that the real parameters \( (\sigma^2, \gamma, \lambda) \) are known such that the backward formula, expressing the transformed jump density in terms of the option prices, is given by
\[
\mathcal{F} \mu(v) = T^{-1} \log(1 - v(v-i) \mathcal{O}(\nu)) + \sigma^2(v-i)^2/2 - i\gamma(v-i) + \lambda, \quad v \in \mathbb{R}.
\]

We construct an empirical version \( \hat{\mathcal{O}} \) from the observations \( (\mathcal{O}_j) \) (e.g. by linear interpolation) and obtain by substitution in this formula an empirical version \( \mathcal{F} \hat{\mu}(v) \) of \( \mathcal{F} \mu(v) \), which satisfies for small noise levels
\[
\|\mathcal{F}(\hat{\mu} - \mu)(v)\| = T^{-1} \| \log \left( \frac{1 - v(v-i) \mathcal{O}(\nu)}{1 - v(v-i) \mathcal{O}(\nu)} \right) \| \approx T^{-1} \left\| \frac{v(v-i) \mathcal{O}(\hat{\mathcal{O}} - \mathcal{O})(v)}{1 - v(v-i) \mathcal{O}(\nu)} \right\|.
\]

This first-order analysis already reveals the ill-posedness of the calibration problem: the higher the frequency \( \|v\| \), the more the error \( \mathcal{F}(\hat{\mathcal{O}} - \mathcal{O})(v) \) in the observation domain is amplified by the factor \( v(v-i) \) as well as by the denominator which tends to zero for \( \|v\| \to \infty \). The nonlinearity is reflected by a noise behavior which depends on the unknown true value \( \mathcal{F} \mathcal{O}(v) \).

A natural approach is to cut-off high frequencies and to consider for \( U > 0 \) the family of estimators
\[
\hat{\mu}_U(x) = \mathcal{F}^{-1}(\mathcal{F} \hat{\mu} \mathbf{1}_{[-U,U]})(x), \quad x \in \mathbb{R}.
\]

Let us write abstractly
\[
\mathcal{F} \hat{\mu}(v) = \mathcal{F} \mu(v) + \sigma(v) \xi(v), \quad v \in \mathbb{R},
\]

with \( \sigma(v) \) denoting the noise behavior and \( \xi(v) \) denoting the normalized noise variable, that is the difference between empirical and true value divided by its standard deviation, at frequency \( v \). Now we see the analogy with the sequence space model (2) analyzed before. In fact, the only difference is the continuity of the spectral parameter \( v \) instead of \( k \in \mathbb{N} \) and the additional difficulty is the much more complicated noise structure. Nevertheless, the estimators \( \hat{\mu}_U \) are provably rate-optimal for the right choice of the cut-off frequency \( U \) (the problem is severely ill-posed in general). In simulations the standard data-driven choices of \( U \) (e.g. Lepski’s method, cross validation) had a poor performance compared to the oracle choice, mostly because of a very badly known noise structure. This is why the quasi-optimality criterion is of interest here.
4.2. Experimental setup

In total we performed 1000 independent experiments. Each of them was set up as follows (notation as in Belomestny and Reiß (2006) where more details can be found):

- 100 design points \((x_i)\) were chosen at random, 50 according to a standard normal and 50 according to a uniform distribution on \([-4, 8]\).
- Corresponding observations \((O_i)\) are generated by calculating the exact value and adding standard normal noise of 3% relative noise level, i.e. \(O_i = O(x_i)(1 + 0.03 e_i)\) with \(e_i \sim \mathcal{N}(0,1)\). The other characteristics are chosen as follows (corresponding to the so-called Merton model): volatility \(\sigma = 0.1\), jumps are standard normal with intensity \(\lambda = 5\) (the jump density \(\nu\) in figure 1 (right)), maturity \(T = 0.25\). The value of \(\gamma\) is obtained by imposing a martingale condition. In figure 1 (left), the true option price function \(O(\cdot)\) is shown together with the observations as a function of \(x\). Abstractly, one can show that in this setting both the bias and the variance term have exponential decay respectively growth in \(U\).
- In total 60 cut-off frequencies \((U_n)\) were chosen with step-width 0.8, i.e. \(U_n = 0.8n\), \(n = 1, \ldots, 60\).

4.3. Parameter choice methods

We will compare three different parameter choice methods. The quasi-optimality and the hardened balancing principle are treated in this paper and Lepski’s method serves as a widely used benchmark.

**Quasi-optimality.** We use exactly as in definition 2.5 with \(\chi(\cdot) = 1\)

\[
n_{qo} := \arg\min_n \{\|\hat{x}^{(n+1)} - \hat{x}^{(n)}\|\}.
\]

**Lepski-type method.** Define

\[
f(n) := \max_{n < m \leq N} \{4^{-1}\|\hat{x}^{(m)} - \hat{x}^{(n)}\|/s(m)\}, \quad n_{\text{Lepski}} := \arg\min_n \{\forall m \geq n : f(m) \leq \kappa\}.
\]
For the choice of $\kappa$ one should theoretically use a quantity depending on the noise behavior and larger than 1. In practice, however, it is observed that $\kappa \in [0.25, 0.75]$ gives superior results and works in more or less any situation. Here we choose $\kappa = 0.75$.

**Hardened balancing principle.** We reuse the computed quantity $f(n)$ defined above and choose as regularization parameter

$$n_{HBP} := \arg\min_n \{f(n)\sqrt{s(n)}\}.$$  

Due to the definition of $f$ it can be seen as a stabilized version of definition 2.5 with $\chi(n) = s(n)^{-1/2}$. This stabilization mainly counters the effects of a subsampling with inappropriately small spacing between the cut-off points.

**Implementation details.** There are certain fine points in the implementation, which are not crucial, but yield slightly superior results. The important point for the methods proposed here is the validity of assumption 2.1. Therefore we estimated $s(\cdot)$ out of 10 independent data sets (each of them constructed as given above, i.e. each of them with different design points and random error) and removed the cut-off parameters $n \in \{17, 27, 36, 44, 46, 48\}$ which seem to violate the assumption empirically.

The Lepski balancing principle and hardened balancing show undesirable properties when $s(\max)/s(n) < 2$. Therefore, we restricted the region of admissible regularization parameters further to $n \in \{1, \ldots, 43\}\setminus\{17, 27, 36\}$. The remaining cut-off points were still used to calculate the values $f(n)$.

### 4.4. Experiment

In figure 2 the distribution of the efficiencies is displayed, i.e. the ratio of errors between the unsupervised solution selected by one of the methods and the optimal one, i.e.,

$$\text{efficiency}(x, y, n_{\text{chosen}}) := \frac{\|x^{(n_{\text{chosen}})} - x\|}{\min_n \|x^{(n)} - x\|}.$$  

Note that the ratio cannot be smaller than 1. Remark further that it is much harder to obtain a low value for $\mathbb{E}$ efficiency $(x, y, n_{\text{chosen}})$ compared to an oracle criterion of the
form $\mathbb{E}\|x^{(\text{optimal})} - x\|/\min_n \mathbb{E}\|x^{(n)} - x\|$ because already one much better estimator in the denominator can spoil the results in our case. The bar plot of figure 2 bins the solutions according to their efficiency. The bins are taken on a double exponential scale, ratios larger than 64 were considered as huge.

As we can see, quasi-optimality and hardened balancing, as defined here, are superior to the Lepski-balancing principle, which suffers from a considerable number of bad results. Although in this situation the hardened balancing principle performs slightly worse, it has a particular advantage. Generally, it is relatively insensitive toward a bad choice of the subsampling whereas already one badly chosen distance in the subsampling could bring quasi-optimality out of track. Yet, it should be recalled that it relies significantly on an estimate of the stochastic noise behavior $s(\cdot)$, obtained from additional data sets.

5. Perspectives

Numerical experiments indicate that the results can be carried over to other regularization methods like Tikhonov regularization. Unfortunately, the correlation structure induced by Tikhonov regularization makes an analysis in comparison to the spectral cut-off regularization much more demanding and remains a topic of future research.

Interestingly, a way to improve the numerical results for spectral cut-off even further is to use a two-step procedure, where first a regularization parameter is obtained from a very coarse subsampling and then the region around the chosen regularization parameter is reconsidered with a finer subsampling. These types of procedures work provably well for change point detection problems (e.g. Korostelev (1987)), a setting, which is to some extent related to our search for the intersection point of $b(\cdot)$ and $s(\cdot)$.

6. Proofs

In order to evaluate the probabilities of staying above or below the threshold, we will make use of the following technical result.

Lemma 6.1. Let $Z = \sum_{k=1}^{\infty} \alpha_k^2 \zeta_k^2$ with $\sum_{k=1}^{\infty} \alpha_k^2 = 1$ and $\zeta_k \sim N(0, 1)$ i.i.d. Then

$\forall z \in (0, 1): \mathbb{P}(Z \leq z) \leq \exp \left( \frac{1 - z + \log(z)}{2 \max_k \alpha_k^2} \right), \quad \forall z > 0: \mathbb{P}(Z \geq z) \leq \sqrt{2} \exp^{-z^{1/4}}$.

Proof. For all $\lambda > 0$ and $z \in (0, 1)$ we have

$\mathbb{P}(Z \leq z) = \mathbb{P} \left( \exp \left( -\lambda \left( \sum_{k=1}^{\infty} \alpha_k^2 \zeta_k^2 \right) \right) \geq e^{-\lambda z} \right)$

$\leq e^{\lambda z} \prod_{k=1}^{\infty} \mathbb{E} \left[ \exp \left( -\lambda \alpha_k^2 \zeta_k^2 \right) \right]$

$= e^{\lambda z} \prod_{k=1}^{\infty} (1 + 2\lambda \alpha_k^2)^{-1/2}$.

Due to $(1 + 2\lambda x)^{-1/2} \leq \exp \left( -\log(1 + x) / \lambda \right)$ for $x \in [1, 1 + \epsilon], \epsilon > 0$, we deduce for $\lambda = \epsilon(2 \max_k \alpha_k^2)^{-1}$ and $\epsilon = z^{-1} - 1$,

$\mathbb{P}(Z \leq z) \leq \exp \left( \frac{\epsilon}{2 \max_k \alpha_k^2} \left( z - \frac{\log(1 + \epsilon)}{\epsilon} \right) \right) = \exp \left( \frac{1 - z + \log(z)}{2 \max_k \alpha_k^2} \right)$.
By Jensen’s inequality we obtain for $z > 0$,
\[
\mathbb{P}(Z \geq z) \leq e^{-z^2/4} \mathbb{E} \left[ \exp \left( \sum_{k=1}^{\infty} \frac{\alpha_k^2 \xi_k^2}{4} \right) \right] \leq e^{-z^2/4} \mathbb{E} [\exp(\xi^2/4)] = \sqrt{2} e^{-z^2/4}. \tag{5}
\]

**Proof of lemma 3.2.** For each $n \geq 1$ and $\alpha > 0$ we derive from the concavity of the log-function and Jensen’s inequality,
\[
R_\alpha(\hat{x}^{(n)})^2 = \mathbb{E}[\|\hat{x}^{(n)} - x\|^2/\alpha] \\
\geq \exp(\mathbb{E}[\log(\|\hat{x}^{(n)} - x\|^2)]) \\
= \exp \left( \log \left( \sum_{k=1}^{\ell(n)} \frac{\sigma(k)^2 \xi_k^2}{\rho(n) + \beta(n)} \right) + \log(\rho(n) + \beta(n)) \right) \\
\geq \exp \left( \sum_{k=1}^{\ell(n)} \frac{\sigma(k)^2 \xi_k^2}{\rho(n) + \beta(n)} \mathbb{E} \left[ \log(\xi_k^2) \right] + \sum_{k=\ell(n)+1}^{\infty} \frac{\gamma(k)^2}{\rho(n) + \beta(n)} \mathbb{E} \left[ \log(\xi_k^2/\gamma(k)^2) \right] \right) (\rho(n) + \beta(n)) \\
= (\rho(n) + \beta(n)) e^{\mathbb{E}[\log(\xi_k^2)]} = e^{\mathbb{E}[\log(\xi_k^2)]} R_2(\hat{x}^{(n)})^2,
\]
which gives the first inequality. For the second, we use $w^{\alpha/2}/\Gamma\left(\frac{\alpha}{2} + 1\right) \leq e^w$ for $w \geq 0$, lemma 6.1 and $(1 - w)^{-1/2} \leq 2w$ for $w \in (0, 1/2]$ to obtain for any $\alpha > 0$ and $\lambda \in (0, \min(\gamma(\ell(n) + 1)^{-2}, \sigma(\ell(n))^{-2})/4),$
\[
\mathbb{E}[\|\hat{x}^{(n)} - x\|^\alpha] = \Gamma\left(\frac{\alpha}{2} + 1\right)\lambda^{-\alpha/2} \mathbb{E}[\|\hat{x}^{(n)} - x\|^\alpha/\Gamma\left(\frac{\alpha}{2} + 1\right)] \\
\leq \Gamma\left(\frac{\alpha}{2} + 1\right)\lambda^{-\alpha/2} \mathbb{E}[\exp(\lambda \|\hat{x}^{(n)} - x\|^\alpha)] \\
= \Gamma\left(\frac{\alpha}{2} + 1\right)\lambda^{-\alpha/2} \exp(2\log(2)\lambda(\rho(n) + \beta(n))).
\]
The choice $\lambda = (\rho(n) + \beta(n))^{-1}$ fulfills the above requirement and thus yields
\[
R_\alpha(\hat{x}^{(n)}) = \mathbb{E}[\|\hat{x}^{(n)} - x\|^\alpha/\alpha] \leq (4\Gamma\left(\frac{\alpha}{2} + 1\right))^{1/\alpha} R_2(\hat{x}^{(n)}).
\]
The last inequality follows from the two others together with the fact that assumption 2.1 implies
\[
\min R_2(\hat{x}^{(n)})^2 \geq \min(b(n^8), s(n^8 + 1)) \geq \frac{C_b - 1}{2} R_2(\hat{x}^{(n^8)})^2.
\]

**Proof of proposition 3.3.** Let us first introduce the following quantities:
\[
a(n) := \sqrt{\chi(n)\sqrt{(s(n+1) - s(n)) + (b(n) - b(n+1))}}, \\
D(n) := \chi(n)\|\hat{x}^{(n+1)} - \hat{x}^{(n)}\|^2, \\
Z(n) := \sqrt{D(n)/a(n)}.
\]
We shall only consider $n > n^8$ because the case $n < n^8$ is symmetric (exchanging $s$ and $b$) and the case $n = n^8$ is trivial. For $m := n - n^8 \geq 1$ we find by assumptions 2.1 and 2.3,
\[
\frac{a(n^8)^2}{a(n^8 + m)^2} \leq \frac{\chi(n^8)}{\chi(n^8 + m)} \frac{(C_s - 1)s(n^8 + m)}{(C_s - 1)s(n^8 + m)} \\
\leq C_\chi m \frac{(C_s - 1)s(n^8 + m)}{(C_s - 1)c^m s(n^8)} \\
\leq C_b C_s - 1 - \left( \frac{c_s}{C_\chi^m} \right) = \rho(n).
\]
Using lemma 6.1, we infer for any $T > 0$,
\[ \tilde{P}(n^* = n) \leq \tilde{P}(D(n) \leq D(n^*)) \leq P(Z(n)^2 \leq \rho(n)Z(n^*)) \leq \tilde{P}(Z(n)^2 \leq T \rho(n) \log(\rho(n)^{-1})) + \tilde{P}(Z(n^*)^2 \geq T \log(\rho(n)^{-1})). \]
Choosing $T = 2r$ we get for the two separate parts,
\[ \tilde{P}(Z(n)^2 \leq T \rho(n) \log(\rho(n)^{-1})) \leq \min \left\{ 1, \exp \left( 1 - \frac{T \rho(n) \log(\rho(n)^{-1})}{2 \max \{ k^{1/p} \}} \right) \right\} \leq \min \left\{ 1, (2re \rho(n) \log(\rho(n)^{-1}))^{r/2} \right\} \leq (2re \rho(n) \log(\rho(n)^{-1}))^{r/2}. \]
In this inequality chain we used that $0 \leq r \leq r(n)$ and that we can distinguish between the case $2re \rho(n) \log(\rho(n)^{-1}) > 1$ (where the minimum above is 1 in any case) and the case $2re \rho(n) \log(\rho(n)^{-1}) \leq 1$.

The second part gets
\[ \tilde{P}(Z(n^*)^2 \geq T \log(\rho(n)^{-1})) \leq \sqrt{2} \exp(-T \log(\rho(n)^{-1})/4) = \sqrt{2} \rho(n)^{r/2} \leq \sqrt{2} (\rho(n) \log(\rho(n)^{-1}))^{r/2} \]
which yields the proposition.

**Proof of theorem 3.4.** To prove that $n^*$ is well defined, we infer from the proof of proposition 3.3 that for all $n > n^*$,
\[ \tilde{P}(D(n) \leq D(n^*)) \leq (\sqrt{2} + (2er)^{r/2} (\rho(n) \log(\rho(n)^{-1}))^{r/2}. \]
This exponential decay implies
\[ \lim_{m \to \infty} \tilde{P}(\exists n \geq m : D(n) \leq D(n^*)) = \lim_{m \to \infty} \sum_{n=m}^{\infty} \tilde{P}(D(n) \leq D(n^*)) = 0, \]
which means that the probability that the criterion $D(\cdot)$ is larger for some $n \geq m$ than at $n^*$ tends to zero as $m \to \infty$, hence $n^* = \arg\min_n D(n)$ is well defined with probability one.

For the main assertion we use H"older’s inequality with $p^{-1} + q^{-1} = 1$, lemma 3.2 and assumption 2.1 for any $\alpha > 0$ to obtain
\[ \tilde{E}[\|\hat{x}(n^*) - x\|^\alpha] = \sum_{m=-n^*+1}^{\infty} \tilde{E}[\|\hat{x}(m) - x\|^\alpha 1_{[m=n^*+m]}] \]
\[ \leq \sum_{m=-n^*+1}^{\infty} \tilde{E}[\|\hat{x}(m) - x\|^\alpha 1_{[m=n^*+m]}]^{1/p} \tilde{P}(n^* = n^* + m)^{1/q} \]
\[ \leq \sum_{m=-n^*+1}^{\infty} R_{\text{exp}}(\hat{x}(m))^{\alpha} \tilde{P}(n^* = n^* + m)^{1/q} \]
\[ \leq (4\Gamma(\frac{np}{T} + 1))^{1/p} \sum_{m=-n^*+1}^{\infty} R_{\text{exp}}(\hat{x}(m))^{\alpha} \tilde{P}(n^* = n^* + m)^{1/q} \]
\[
\begin{align*}
\lesssim (4\Gamma(\frac{q\alpha}{2} + 1))^{1/p} (2s(n^\#))^{\alpha/2} & \left(\sum_{m=0}^{\infty} C^{am/2}_s \hat{p}(n^* = n^\# + m)^{1/q}\right) \\
& + \sum_{m=0}^{n^\#-1} C^{am/2}_b \hat{p}(n^* = n^\# - m)^{1/q}.
\end{align*}
\]

We choose \(c_q > 0\) such that
\[
\alpha \leq \frac{r}{\max \left\{ \log(C_s), \log(C_b) \right\} (1 + 3c_q)},
\]
\[
q := \alpha \max \left\{ \frac{\log(C_s)}{\log(C_s/C_b)}, \frac{\log(C_b)}{\log(C_b/C_s)} \right\} (1 + 2c_q) > 1
\]
and infer from proposition 3.3,
\[
\sum_{m=1}^{\infty} C^{am/2}_s \hat{p}(n^* = n^\# + m)^{1/q} \leq \left(\sqrt{2} + (2r)^{q/2}\right)^{1/q} \left(\frac{C_b C_s - 1}{c_s - 1} \log \left(\frac{c_s - 1}{C_b C_s - 1} C_s \right)\right)^{r/2q}
\times \sum_{m=1}^{\infty} C^{am/2}_s \left(\frac{C_s}{C_X}\right)^{-rm/2q} m^{1/2q}.
\]

Looking just at the sum we get because of \(c_s > C_s\) and \(\max \left\{ \log(C_s), \log(C_b) \right\} > 0,
\[
\sum_{m=1}^{\infty} C^{am/2}_s \left(\frac{C_s}{C_X}\right)^{-rm/2q} m^{1/2q} \leq \sum_{m=1}^{\infty} \left(\frac{C_s}{C_X}\right)^{-amc_q} \max \left\{ \frac{\log(C_s)}{\log(C_s/C_b)}, \frac{\log(C_b)}{\log(C_b/C_s)} \right\} m^{oc_q} \max \left\{ \frac{\log(C_s)}{\log(C_s/C_b)}, \frac{\log(C_b)}{\log(C_b/C_s)} \right\}
\leq K_1(c_s, c_b, C_s, C_b, \alpha, c_q) \leq \infty.
\]
Equivalently we get
\[
\sum_{m=0}^{n^\#-1} C^{am/2}_b \hat{p}(n^* = n^\# - m)^{1/q} \leq K_2(c_s, c_b, C_s, C_b, \alpha, c_q) < \infty.
\]
Hence we conclude that \(\tilde{E}[\|\hat{\xi}(n^\#) - x\|^\alpha]\) is bounded by a multiple of \(s(n^\#)^{\alpha/2}\), which has the order of \(R_2(\hat{\xi}(n^\#))^\alpha\). We eventually obtain with constants \(K = K(c_s, c_b, C_s, C_b, r, \alpha), K' = K'(c_s, c_b, C_s, C_b, r, \alpha)\) (note that \(p, q\) and \(c_q\) depend on \(r, \alpha\) and the remaining constants),
\[
R_q(\hat{\xi}(n^\#)) \leq K' R_2(\hat{\xi}(n^\#)) \leq K \min_{n \geq 1} R_q(\hat{\xi}(n^\#)).
\]

\[\square\]

**Remark 6.2.** As we have seen, the factors \(K_1(c_s, c_b, C_s, C_b, \alpha, c_q), K_2(c_s, c_b, C_s, C_b, \alpha, c_q), K'(c_s, c_b, C_s, C_b, r, \alpha)\) and \(K(c_s, c_b, C_s, C_b, r, \alpha)\) are independent of the intersection point \(n^\#\) such that we can treat them as constants in the oracle inequality.

**Acknowledgments**

We are grateful for the constructive criticism by two referees and a member of the editorial board. The first author gratefully acknowledges the financial support by the Upper Austrian Technology and Research Promotion.
References

Bakushinskii A 1984 Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion *Comput. Math. Math. Phys.* 24 181–2

Bauer F 2007 Some considerations concerning regularization and parameter choice algorithms *Inverse Problems* 23 837–58

Bauer F and Pereverzev S 2005 Regularization without preliminary knowledge of smoothness and error behavior *Eur. J. Appl. Math.* 16 303–17

Belomestny D and Reiß M 2006 Spectral calibration of exponential Lévy models *Finance Stoch.* 10 449–74

Cohen A, Hoffmann M and Reiß M 2004 Adaptive wavelet Galerkin methods for linear inverse problems *SIAM J. Numer. Anal.* 42 1479–501

Cont R and Tankov P 2004 Financial Modelling with Jump Processes (Chapman and Hall/CRC Financial Mathematics Series) (Boca Raton, FL: CRC Press)

Crépey S 2003 Calibration of the local volatility in a generalized Black–Scholes model using Tikhonov regularization *SIAM J. Math. Anal.* 34 1183–206

Egger H, Hein T and Hofmann B 2006 On decoupling of volatility smile and term structure in inverse option pricing *Inverse Problems* 22 1247–59

Engl H, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer)

Kaipio J and Somersalo E 2005 *Statistical and Computational Inverse Problems* (Applied Mathematical Sciences vol 160) (New York: Springer)

Korostelev A 1987 On minimax estimation of a discontinuous signal *Theory Probab. Appl.* 32 727–30

Lepski O 1990 On a problem of adaptive estimation in Gaussian white noise *Theory Probab. Appl.* 35 454–66

Tikhonov A and Arsenin V 1977 *Solutions of Ill-Posed Problems* (New York: Wiley)

Tikhonov A, Glasko V and Kriksin Y 1979 On the question of quasi-optimal choice of a regularized approximation *Sov. Math. Dokl.* 20 1036–40

Wahba G 1977 Practical approximate solutions to linear operator equations when the data are noisy *SIAM J. Numer. Anal.* 14 651–67

Wasserman L 2006 *All of Nonparametric Statistics, Springer Texts in Statistics* (New York: Springer)