# MULTI-PARAMETER TIKHONOV REGULARIZATION*

KAZUFUMI ITO$^\dagger$, BANGTI JIN$^\ddagger$, AND TOMOYA TAKEUCHI$^\S$

**Abstract.** We study multi-parameter Tikhonov regularization, i.e., with multiple penalties. Such models are useful when the sought-for solution exhibits several distinct features simultaneously. Two choice rules, i.e., discrepancy principle and balancing principle, are studied for choosing an appropriate (vector-valued) regularization parameter, and some theoretical results are presented. In particular, the consistency of the discrepancy principle as well as convergence rate are established, and an a posteriori error estimate for the balancing principle is established. Also two fixed point algorithms are proposed for computing the regularization parameter by the latter rule. Numerical results for several nonsmooth multi-parameter models are presented, which show clearly their superior performance over their single-parameter counterparts.

**Key words.** Multi-parameter regularization, value function, balancing principle, parameter choice.

**AMS subject classifications.** 65J20, 65J22.

**1. Introduction.** In this paper, we are interested in solving linear inverse problems

$$Kx = y^\delta, \tag{1}$$

where $y^\delta \in Y$ is a noisy version of the exact data $y^\dagger = Kx^\dagger \in Y$ with $\delta^2 = \phi(x^\dagger, y^\delta)$ being the noise level, the operator $K : X \to Y$ is bounded and linear, and the spaces $X$ and $Y$ are Banach spaces.

Typically, problem (1) suffers from ill-posedness in the sense that a small perturbation in the data might lead to large deviations in the retrieved solution, and this often poses great challenges to their stable yet accurate numerical solution. Usually, a regularization strategy is applied to find a stable approximate solution [18, 6]. The most widely adopted approach is Tikhonov regularization, which seeks an approximation $x^\delta_{\boldsymbol\eta}$ to problem (1) by minimizing the following Tikhonov functional

$$J_{\boldsymbol\eta}(x) = \phi(x, y^\delta) + \boldsymbol\eta \cdot \boldsymbol\psi(x).$$

Here the functionals $\phi$ and $\boldsymbol\psi$ represent data fidelity and (vector-valued) penalty, respectively, and $\boldsymbol\eta \cdot \boldsymbol\psi(x)$ is the dot product between $\boldsymbol\eta = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$ and $\boldsymbol\psi(x) = (\psi_1(x), \ldots, \psi_n(x))^{\mathrm{T}}$, i.e., $\boldsymbol\eta \cdot \boldsymbol\psi(x) = \sum_{i=1}^n \eta_i \psi_i(x)$. Common choices of the fidelity $\phi(x, y^\delta)$ include $\|Kx - y^\delta\|_{L^2}^2$, $\|Kx - y^\delta\|_{L^1}$ and $\int(Kx - y^\delta \ln Kx)$, which are statistically well suited to additive Gaussian noise, Laplace (impulsive) noise and Poisson noise, respectively. The penalties $\psi_i$ are nonnegative, convex and (weak) lower semicontinuous. The typical choice includes $\|x\|_{L^2}^2$, $\|x\|_{\ell^p}^p$, $\|x\|_{H^m}^2$ and $|x|_{TV}$ etc. The regularization parameter vector $\boldsymbol\eta$ compromises fidelity with penalties.

The use of multiple penalties, henceforth called multi-parameter regularization, in the functional $J_{\boldsymbol\eta}$ is motivated by practical applications which exhibit multiple/multiscale features. We just take microarray data analysis for an example. Here

the number of data is often far less than that of the unknowns. A desirable approach should select all variables relevant to the proper functioning of gene network. The conventional $\ell^2$ penalty tends to select all variables, including irrelevant ones, since the resulting estimate has almost no nonzero entries. To remedy this issue, $\ell^1$ penalty has been suggested as an alternative. However, the $\ell^1$ approach delivers undesirable results for problems where there are highly correlated features and all relevant ones are to be identified in that it tends to select only one feature out of the relevant group instead of all relevant features of the group [20], thereby missing the groupwise structure. Zou and Hastie [20] proposed the elastic-net by incorporating the $\ell^2$ penalty into the $\ell^1$ penalty, in the hope of retrieving the whole relevant group, and numerically demonstrated its excellent performance for simulation studies and real-data applications. Such multiple/multiscale features appear also in many other applications, e.g., image processing [14, 17], electrocardiography [3], and geodesy [19].

A number of experimental studies [3, 20, 19] have shown great potential of multi-parameter models for better capturing multiple distinct features of the solution. However, a general theory for such models remains largely under-explored. There are several attempts on various aspects, e.g., parameter choice, convergence and statistical interpretation [1, 2, 4, 9, 5, 12, 13] of multi-parameter regularization. For instance, Lu et al [12] discussed the discrepancy principle using Hilbert space scales, and derived some error estimates, but the parameter is vastly nonunique and it remains unclear which one to use. They also adapted the model function approach to choose the regularization parameter, but the underlying mechanism remains unclear. Jin et al [9] recently investigated the properties, e.g., consistency and error estimates, of elastic-net for asymptotically linear coupling between the two terms, and proposed two active-set type methods for efficient numerical realization.

This paper aims at developing some theory for such models in a general framework. The value function and its properties are first derived. Then two parameter choice rules, i.e., discrepancy principle and balancing principle, are studied. The consistency and convergence rates are established for the former. The balancing principle can be derived from the Bayesian inference [10], and it was generalized in [8]. The principle balances the penalty with the fidelity term. The variant under consideration here is solely based on the value function, and does not require a knowledge of the noise level. An a posteriori error estimate is derived, and two efficient numerical algorithms are also proposed.

The rest of the paper is structured as follows. In Section 2, we investigate the value function and derive some properties, e.g., monotonicity, concavity, asymptotic and especially differentiability. In Section 3, we investigate two parameter choice rules, i.e., discrepancy principle and balancing principle, and discuss their theoretical properties. In addition, two fixed point algorithms for the efficient numerical realization of the balancing principle are proposed. Numerical results for several examples are presented in Section 4 to illustrate the efficiency and accuracy of the proposed approaches. Finally, we conclude the paper with several future research topics.

*Notation.* Let $x_{\boldsymbol{\eta}}^{\delta}$ be a minimizer to the functional $J_{\boldsymbol{\eta}}(x)$, and $\mathcal{M}_{\boldsymbol{\eta}}$ be the set of minimizers. For vectors $\boldsymbol{\eta} \in \mathbb{R}^n$ and $\hat{\boldsymbol{\eta}} \in \mathbb{R}^n$, we denote by $\boldsymbol{\eta} \leq \hat{\boldsymbol{\eta}}$ if $\eta_i \leq \hat{\eta}_i \, \forall 1 \leq i \leq n$.

**2. The value function and its properties.** In this section, we collect important properties of the value function $F(\boldsymbol{\eta})$ defined by

$$(2) \qquad\qquad F(\boldsymbol{\eta}) = \inf_{x \in Q_{\mathrm{ad}}} J_{\boldsymbol{\eta}}(x),$$

where the set $Q_{\mathrm{ad}}$ stands for a convex constraint. Here, the existence of a minimizer $x_{\boldsymbol{\eta}}^{\delta}$ to the functional $J_{\boldsymbol{\eta}}$ is not a priori assumed. Provided that a minimizer $x_{\boldsymbol{\eta}}^{\delta}$ does exist, we have $F(\boldsymbol{\eta}) = J_{\boldsymbol{\eta}}(x_{\boldsymbol{\eta}}^{\delta})$. The value function $F$ will play an important role in developing a balancing principle, see Section 3.2. The results presented below generalize those for the single parameter [7], and the proofs are similar and thus omitted.

A first result shows the continuity and concavity of $F$.

LEMMA 2.1. *The value function $F(\boldsymbol{\eta})$ is monotonically increasing in the sense $F(\hat{\boldsymbol{\eta}}) \leq F(\boldsymbol{\eta})$ if $\hat{\boldsymbol{\eta}} \leq \boldsymbol{\eta}$, and it is concave.*

REMARK 2.1. *Lemma 2.1 does not require the existence of $x \in Q_{\mathrm{ad}}$ that achieves the infimum of $J_{\boldsymbol{\eta}}$. The results are also true for nonlinear operators and in the presence of convex constraint $Q_{\mathrm{ad}}$.*

Next we examine the properties of the value function $F$ more closely. Recall first one-sided partial derivatives $\partial_i^{\pm} F$ are defined by

$$\partial_i^- F(\boldsymbol{\eta}) = \lim_{h \to 0^+} \frac{F(\boldsymbol{\eta}) - F(\boldsymbol{\eta} - h\boldsymbol{e}_i)}{h},$$

$$\partial_i^+ F(\boldsymbol{\eta}) = \lim_{h \to 0^+} \frac{F(\boldsymbol{\eta} + h\boldsymbol{e}_i) - F(\boldsymbol{\eta})}{h},$$

where $\boldsymbol{e}_i$ is the $i$th canonical basis.

The next result shows some properties, i.e., existence, nonnegativity, monotonicity and (left- and right-) continuity, of the one-sided partial derivatives $\partial_i^{\pm}$. The properties follow directly from Lemma 2.1.

LEMMA 2.2. *For any $\boldsymbol{\eta} > 0$, there hold*
  (i) *The one-sided partial derivatives $\partial_i^{\pm} F(\boldsymbol{\eta})$ exist, and $\partial_i^{\pm} F(\boldsymbol{\eta}) \geq 0$;*
 (ii) *For any $h > 0$, there holds $0 \leq \partial_i^+ F(\boldsymbol{\eta} + h\boldsymbol{e}_i) \leq \partial_i^- F(\boldsymbol{\eta} + h\boldsymbol{e}_i) \leq \partial_i^+ F(\boldsymbol{\eta}) \leq \partial_i^- F(\boldsymbol{\eta})$;*
(iii) *$\partial_i^- F(\boldsymbol{\eta}) = \lim_{h \to 0^+} \partial_i^- F(\boldsymbol{\eta} - h\boldsymbol{e}_i)$ and $\partial_i^+ F(\boldsymbol{\eta}) = \lim_{h \to 0^+} \partial_i^+ F(\boldsymbol{\eta} + h\boldsymbol{e}_i)$.*

REMARK 2.2. *The partial differentiability of $F$ in the $i$-th direction at $\boldsymbol{\eta}$ guarantees the continuity of $\partial_i^{\pm} F$ at this point. Indeed, the monotonicity of $\partial_i^{\pm} F$ and the left continuity of $\partial_i^+ F$ yield the inequalities*

$$\partial_i^+ F(\boldsymbol{\eta}) = \lim_{h \to 0^+} \partial_i^+ F(\boldsymbol{\eta} + h\boldsymbol{e}_i) \leq \lim_{h \to 0^+} \partial_i^- F(\boldsymbol{\eta} + h\boldsymbol{e}_i) \leq \partial_i^- F(\boldsymbol{\eta}).$$

*Now suppose $F$ is differentiable at $\boldsymbol{\eta}$, i.e., $\partial_i^- F(\boldsymbol{\eta}) = \partial_i^+ F(\boldsymbol{\eta})$. Then from the inequalities it follows that*

$$\lim_{h \to 0^+} \partial_i^- F(\boldsymbol{\eta} + h\boldsymbol{e}_i) = \partial_i^- F(\boldsymbol{\eta}),$$

*which shows the continuity of $\partial_i^- F$ at $\boldsymbol{\eta}$. Similarly it follows that $\partial_i^+ F$ is continuous at $\boldsymbol{\eta}$.*

The asymptotic behavior of $F(\boldsymbol{\eta})$ is useful for designing numerical algorithms.

PROPOSITION 2.1. *The following asymptotics of $F$ hold*

$$\lim_{|\boldsymbol{\eta}| \to 0} F(\boldsymbol{\eta}) = \inf_{x \in Q_{ad}} \phi(x, y^{\delta}) \quad and \quad \lim_{|\boldsymbol{\eta}| \to 0} \eta_i \partial_i^{\pm} F(\boldsymbol{\eta}) = 0.$$

The partial derivatives $\partial_i^\pm F$ are closely connected to the fidelity $\phi$ and penalty $\psi$ under the assumption of existence of a minimizer, i.e., the set $\mathcal{M}_{\boldsymbol\eta}$ is nonempty. This is guaranteed by:

ASSUMPTION 2.1. *The functionals $\phi$ and $\psi_i$ satisfy:*
(i) *For any sequence $\{x_n\}_n \subset Q_{\mathrm{ad}}$ such that $\phi$ and $\psi_i$ for all $1 \le i \le n$ are uniformly bounded, there exists a subsequence $\{x_{n_k}\}_k$ which converges to an element $x^* \in Q_{ad}$ in the $\tau$-topology.*
(ii) *$\phi$ and $\psi_i$ are lower semi-continuous with respect to $\tau$-convergent sequences, i.e., if a subsequence $\{x_n\}_n$ converges to $x^* \in Q_{\mathrm{ad}}$ in $\tau$-topology, then*

$$\phi(x^*) \le \liminf_{n\to\infty} \phi(x_n) \ \ and \ \ \psi_i(x^*) \le \liminf_{n\to\infty} \psi_i(x_n).$$

In case that the set $\mathcal{M}_{\boldsymbol\eta}$ contains multiple elements, there might exist distinct $x_{\boldsymbol\eta}^\delta, \hat{x}_{\boldsymbol\eta}^\delta \in \mathcal{M}_\eta$ such that

$$F(\eta) = \phi(x_{\boldsymbol\eta}^\delta, y^\delta) + {\boldsymbol\eta} \cdot {\boldsymbol\psi}(x_{\boldsymbol\eta}^\delta) = \phi(\hat{x}_{\boldsymbol\eta}^\delta, y^\delta) + {\boldsymbol\eta} \cdot {\boldsymbol\psi}(\hat{x}_{\boldsymbol\eta}^\delta) \quad \text{but} \quad \phi(x_{\boldsymbol\eta}^\delta, y^\delta) \neq \phi(\hat{x}_{\boldsymbol\eta}^\delta, y^\delta),$$

i.e., the functions $\phi(x_{\boldsymbol\eta}^\delta, y^\delta)$ and ${\boldsymbol\psi}(x_{\boldsymbol\eta}^\delta)$ are potentially multi-valued in ${\boldsymbol\eta}$.

A first relation between $\psi_i$ and $\partial_i^\pm F$ is given by

LEMMA 2.3. *Let Assumption 2.1 be fulfilled. Then for any $x_{\boldsymbol\eta}^\delta \in \mathcal{M}_{\boldsymbol\eta}$, there hold*

$$\partial_i^+ F({\boldsymbol\eta}) \le \psi_i(x_{\boldsymbol\eta}^\delta) \le \partial_i^- F({\boldsymbol\eta}) \quad i = 1, \dots, n,$$

$$F({\boldsymbol\eta}) - \sum_{i=1}^n \eta_i \partial_i^- F({\boldsymbol\eta}) \le \phi(x_{\boldsymbol\eta}^\delta, y^\delta) \le F({\boldsymbol\eta}) - \sum_{i=1}^n \eta_i \partial_i^+ F({\boldsymbol\eta}).$$

An immediate consequence of Lemma 2.3 is:

COROLLARY 2.1. *Let Assumption 2.1 be fulfilled. If $\partial_i F({\boldsymbol\eta})$ exists at ${\boldsymbol\eta}$ for all $i$, then $\psi_i(x_{\boldsymbol\eta}^\delta)$ and $\phi(x_{\boldsymbol\eta}^\delta, y^\delta)$ are single valued at ${\boldsymbol\eta}$ and*

$$\partial_i F({\boldsymbol\eta}) = \psi_i(x_{\boldsymbol\eta}^\delta) \quad and \quad F({\boldsymbol\eta}) - {\boldsymbol\eta} \cdot \partial F({\boldsymbol\eta}) = \phi(x_{\boldsymbol\eta}^\delta, y^\delta) \quad \forall x_{\boldsymbol\eta}^\delta \in \mathcal{M}_{\boldsymbol\eta}.$$

More precisely, the partial derivatives $\partial_i^\pm$ can be expressed by $\psi_i$ as follows.

THEOREM 2.1. *Let Assumption 2.1 hold. Then for any ${\boldsymbol\eta} > 0$ and every $i$, there exist $x_i^+, x_i^- \in \mathcal{M}_{\boldsymbol\eta}$ such that*

$$\psi_i(x_i^+) = \partial_i^+ F({\boldsymbol\eta}) \quad and \quad \psi_i(x_i^-) = \partial_i^- F({\boldsymbol\eta}).$$

Theorem 2.1 in conjunction with Lemma 2.3 implies the following corollary.

COROLLARY 2.2. *Let Assumption 2.1 hold. Then*
(i) *There exist $x_i^+, x_i^- \in \mathcal{M}_{\boldsymbol\eta}$ such that $\psi_i(x_i^+) = \min\limits_{x \in \mathcal{M}_{\boldsymbol\eta}} \psi_i(x)$ and $\psi_i(x_i^-) = \max\limits_{x \in \mathcal{M}_{\boldsymbol\eta}} \psi_i(x)$.*
(ii) *If $\psi_i(x_{\boldsymbol\eta}^\delta) = \psi_i(\hat{x}_{\boldsymbol\eta}^\delta)$ for all $x_{\boldsymbol\eta}^\delta, \hat{x}_{\boldsymbol\eta}^\delta \in \mathcal{M}_{\boldsymbol\eta}$ for all ${\boldsymbol\eta} > 0$, then $\partial_i F({\boldsymbol\eta})$ exists and it is continuous.*

The last result gives a sufficient condition for the differentiability of the value function $F$. It plays an important role in especially designing an efficient algorithm for certain choice rules, by e.g., Morozov's principle and balancing principle [11].

THEOREM 2.2. *Assume that the minimizer of the functional $J_{\boldsymbol{\eta}}$ is unique at $\boldsymbol{\eta} > 0$. Then the derivatives $\{\partial_i F(\boldsymbol{\eta})\}_i$ exist and are continuous at $\boldsymbol{\eta}$. In particular, $F$ is differentiable at $\boldsymbol{\eta}$.*

**3. Parameter choice rules.** In this section, we discuss two choice rules, i.e., discrepancy principle [16, 10] and balancing principle [7], for multi-parameter models. For notational simplicity, we shall restrict our attention to the case of two penalty terms.

**3.1. Discrepancy principle.** Here we investigate the discrepancy principle due to Morozov [16] for multi-parameter regularization. We shall assume a triangle-type inequality for the functional $\phi$.

ASSUMPTION 3.1. *The functional $\phi(x, y)$ vanishes if and only if $Kx = y$, and satisfies an inequality $\phi(x, y) \leq c(\phi(x', y') + \phi(x, y'))$ for some constant $c$ and any $x'$ with $Kx' = y$.*

The discrepancy principle determines an appropriate (vector-valued) regularization parameter $\boldsymbol{\eta}$ by

$$(3) \qquad \phi(x_{\boldsymbol{\eta}}^{\delta}, y^{\delta}) = c_m \delta^2$$

for some constant $c_m \geq 1$. The rationale of the principle is that the solution accuracy in terms of the residual should be compatible with the data accuracy (noise level).

THEOREM 3.1. *Let Assumptions 2.1 and 3.1 be satisfied and the operator $K$ be injective. Then for any $\boldsymbol{\eta} \equiv \boldsymbol{\eta}(\delta)$ satisfying (3) and $c_0 \leq \frac{\eta_1(\delta)}{\eta_2(\delta)} \leq c_1$ for some $c_0, c_1 > 0$, there holds $\lim_{\delta \to 0} x_{\boldsymbol{\eta}}^{\delta} = x^{\dagger}$ in $\tau$-topology.*

*Proof.* The minimizing property of $x_{\boldsymbol{\eta}}^{\delta}$ implies

$$\phi(x_{\boldsymbol{\eta}}^{\delta}, y^{\delta}) + \boldsymbol{\eta} \cdot \boldsymbol{\psi}(x_{\boldsymbol{\eta}}^{\delta}) \leq \phi(x^{\dagger}, y^{\delta}) + \boldsymbol{\eta} \cdot \boldsymbol{\psi}(x^{\dagger})$$
$$\leq \delta^2 + \boldsymbol{\eta} \cdot \boldsymbol{\psi}(x^{\dagger}).$$

From the discrepancy equation (3), we deduce

$$(4) \qquad \boldsymbol{\eta} \cdot \boldsymbol{\psi}(x_{\boldsymbol{\eta}}^{\delta}) \leq \boldsymbol{\eta} \cdot \boldsymbol{\psi}(x^{\dagger}).$$

Therefore, either $\psi_1(x_{\boldsymbol{\eta}}^{\delta}) \leq \psi_1(x^{\dagger})$ or $\psi_2(x_{\boldsymbol{\eta}}^{\delta}) \leq \psi_2(x^{\dagger})$ holds. Now the assumption $c_0 \leq \frac{\eta_1(\delta)}{\eta_2(\delta)} \leq c_1$ implies that the sequence $\{\psi_i(x_{\boldsymbol{\eta}}^{\delta}), i = 1, 2\}_{\delta}$ is uniformly bounded. Hence the coercivity of the functional indicates that the sequence $\{x_{\boldsymbol{\eta}}^{\delta}\}_{\delta}$ is uniformly bounded. Thus there exists a subsequence, also denoted by $\{x_{\boldsymbol{\eta}}^{\delta}\}_{\delta}$, and some $x^*$, such that

$$x_{\boldsymbol{\eta}}^{\delta} \to x^* \quad \text{in } \tau\text{-topology.}$$

The $\tau$-lower semicontinuity of the functional $\phi$ and Assumption 3.1 yields

$$0 \leq \phi(x^*, y^{\dagger}) \leq c \liminf_{\delta \to 0}(\phi(x^{\dagger}, y^{\delta}) + \phi(x_{\boldsymbol{\eta}}^{\delta}, y^{\delta})) \leq \liminf_{\delta \to 0} c(1 + c_m)\delta^2 = 0.$$

In particular, $\phi(x^*, y^\dagger) = 0$, i.e., $Kx^* = y^\dagger$. This together with the injectivity $K$ implies $x^* = x^\dagger$. Since every subsequence has a subsubsequence converging to $x^\dagger$, the whole sequence converges to $x^\dagger$. $\square$

REMARK 3.1. *The condition* $c_0 \leq \frac{\eta_1(\delta)}{\eta_2(\delta)} \leq c_1$ *ensures the uniform boundedness of both penalties, and thus we can utilize the lower-semicontinuity of the functionals to arrive at the desired $\tau$-convergence.*

THEOREM 3.2. *Let Assumptions 2.1 and 3.1 hold. If a subsequence $\{\boldsymbol{\eta}(\delta)\}_\delta$ converges and $\tilde{\eta} \equiv \lim\limits_{\delta \to 0} \frac{\eta_1(\delta)}{\eta_2(\delta)} > 0$. Then it contains a subsequence $\tau$-converging to an $\tilde{\eta}\psi_1 + \psi_2$-minimizing solution and*

$$\lim_{\delta \to 0} \left( \frac{\eta_1(\delta)}{\eta_2(\delta)} \psi_1(x_{\boldsymbol{\eta}}^\delta) + \psi_2(x_{\boldsymbol{\eta}}^\delta) \right) = \tilde{\eta}\psi_1(x^\dagger) + \psi_2(x^\dagger).$$

*Moreover, if the $\tilde{\eta}\psi_1 + \psi_2$-minimizing solution is unique, then the whole subsequence $\tau$-converges.*

*Proof.* By repeating the arguments in Theorem 3.1, we deduce that there exists a subsequence, also denoted by $\{x_{\boldsymbol{\eta}}^\delta\}_\delta$, and some $x^*$, such that

$$x_{\boldsymbol{\eta}}^\delta \to x^* \quad \text{in } \tau\text{-topology.}$$

and by the $\tau$-lower-semicontinuity, we have $\phi(x^*, y^\dagger) = 0$. By virtue of lower semi-continuity of the functionals and inequality (4), we deduce

$$\begin{aligned}
\tilde{\eta}\psi_1(x^*) + \psi_2(x^*) &\leq \liminf_{\delta \to 0} \left( \frac{\eta_1(\delta)}{\eta_2(\delta)} \psi_1(x_{\boldsymbol{\eta}}^\delta) + \psi_2(x_{\boldsymbol{\eta}}^\delta) \right) \\
&\leq \limsup_{\delta \to 0} \left( \frac{\eta_1(\delta)}{\eta_2(\delta)} \psi_1(x_{\boldsymbol{\eta}}^\delta) + \psi_2(x_{\boldsymbol{\eta}}^\delta) \right) \\
&\leq \lim_{\delta \to 0} \left( \frac{\eta_1(\delta)}{\eta_2(\delta)} \psi_1(x^\dagger) + \psi_2(x^\dagger) \right) = \tilde{\eta}\psi_1(x^\dagger) + \psi_2(x^\dagger),
\end{aligned}$$

which together with the identity $\phi(x^*, y^\dagger) = 0$ implies that $x^*$ is an $\tilde{\eta}\psi_1 + \psi_2$-minimizing solution. The desired identity follows from the above inequalities with $x^*$ in place of $x^\dagger$. The whole sequence convergence follows from the standard subsequence argument. $\square$

In Theorems 3.1 and 3.2, we have assumed the existence of a solution $\boldsymbol{\eta}$ to equation (3). This is guaranteed if the Tikhonov functional $J_{\boldsymbol{\eta}}$ has a unique minimizer, see Theorem 2.2.

THEOREM 3.3. *Assume that $J_{\boldsymbol{\eta}}$ has a unique minimizer for all $\boldsymbol{\eta} > 0$, $\lim_{|\boldsymbol{\eta}| \to 0} \phi(x_{\boldsymbol{\eta}}^\delta, y^\delta) < c_m \delta^2$, and there is a sequence $\{\boldsymbol{\eta}_n\}$ such that $\lim_{n \to \infty} \phi(x_{\boldsymbol{\eta}_n}^\delta, y^\delta) > c_m \delta^2$. Then there exists at least one solution to (3).*

*Proof.* By Theorem 2.2 and Lemma 2.1, the uniqueness of a minimizer to $J_{\boldsymbol{\eta}}$ for all $\boldsymbol{\eta} > 0$ implies that the function $\phi(x_{\boldsymbol{\eta}}^\delta, y^\delta)$ is continuous in $\boldsymbol{\eta}$. The desired assertion follows from the continuity directly. $\square$

Lastly, we present an error estimate in case of $Y$ being a Hilbert space and $\phi(x, y^\delta) = \|Kx - y^\delta\|^2$ and convex penalties $\boldsymbol{\psi}$. We use Bregman distance to measure the error. Denote the subdifferential of a functional $\psi(x)$ at $x^\dagger$ by $\partial\psi(x^\dagger)$, i.e.,

$\partial\psi(x^\dagger) = \{\xi \in X^* : \psi(x) \geq \psi(x^\dagger) + \langle \xi, x - x^\dagger \rangle \, \forall x \in X\}$, and the Bregman distance $d_\xi(x, x^\dagger)$ by for any $\xi \in \partial\psi(x^\dagger)$

$$d_\xi(x, x^\dagger) := \psi(x) - \psi(x^\dagger) - \langle \xi, x - x^\dagger \rangle.$$

THEOREM 3.4. *If $Y$ is a Hilbert space and the exact solution $x^\dagger$ satisfies the source condition:* $\mathrm{range}(K^*) \cap \partial\psi_1(x^\dagger) \cap \partial\psi_2(x^\dagger) \neq \emptyset$. *Then for any $\boldsymbol{\eta}^*$ solving (3), there exists some $i$ and $\xi_i \in \partial\psi_i(x^\dagger)$ such that*

$$d_{\xi_i}(x^\delta_{\boldsymbol{\eta}^*}, x^\dagger) \leq C\delta.$$

*Proof.* By the minimizing property of $x^\delta_{\boldsymbol{\eta}^*}$, we have

$$\phi(x^\delta_{\boldsymbol{\eta}^*}, y^\delta) + \boldsymbol{\eta}^* \cdot \boldsymbol{\psi}(x^\delta_{\boldsymbol{\eta}^*}) \leq \phi(x^\dagger, y^\delta) + \boldsymbol{\eta}^* \cdot \boldsymbol{\psi}(x^\dagger) \leq \delta^2 + \boldsymbol{\eta}^* \cdot \boldsymbol{\psi}(x^\dagger).$$

The definition of the discrepancy principle indicates

$$\boldsymbol{\eta}^* \cdot \boldsymbol{\psi}(x^\delta_{\boldsymbol{\eta}^*}) \leq \boldsymbol{\eta}^* \cdot \boldsymbol{\psi}(x^\dagger).$$

Consequently, we have that there holds $\psi_i(x^\delta_{\boldsymbol{\eta}^*}) \leq \psi_i(x^\dagger)$ for either $i = 1$ or $i = 2$. Therefore, by the source condition, for some $\xi_i \in \mathrm{range}(K^*) \cap \partial\psi_i(x^\dagger)$ or equivalently $\xi_i = K^* w_i$ for some source representer $w_i$, and the Cauchy-Schwarz inequality, we deduce

$$\begin{aligned} d_{\xi_i}(x^\delta_{\boldsymbol{\eta}^*}, x^\dagger) &= \psi_i(x^\delta_{\boldsymbol{\eta}^*}) - \psi_i(x^\dagger) - \langle \xi_i, x^\delta_{\boldsymbol{\eta}^*} - x^\dagger \rangle \leq -\langle \xi_i, x^\delta_{\boldsymbol{\eta}^*} - x^\dagger \rangle \\ &= -\langle K^* w_i, x^\delta_{\boldsymbol{\eta}^*} - x^\dagger \rangle = -\langle w_i, K(x^\delta_{\boldsymbol{\eta}^*} - x^\dagger) \rangle \\ &\leq \|w_i\| \|K(x^\delta_{\boldsymbol{\eta}^*} - x^\dagger)\| \\ &\leq \|w_i\| \left( \|Kx^\delta_{\boldsymbol{\eta}^*} - y^\delta\| + \|y^\delta - Kx^\dagger\| \right) \leq (1 + c_m)\|w_i\|\delta. \end{aligned}$$

This shows the desired estimate. □

The source condition in Theorem 3.4 can be hard to argue. Alternatively, we can have another convergence rates result under a seemingly less restrictive assumption.

THEOREM 3.5. *If $Y$ is a Hilbert space and the exact solution $x^\dagger$ satisfies the source condition: for any $t \in [0,1]$, there exists $w_t$ such that $K^* w_t = \xi_t \in \partial(t\psi_1(x^\dagger) + (1-t)\psi_2(x^\dagger))$. Then for any $\boldsymbol{\eta}^*$ solving (3), and letting $t^* = \frac{\eta_1^*(\delta)}{\eta_1^*(\delta) + \eta_2^*(\delta)}$, the following estimate holds*

$$d_{\xi_{t^*}}(x^\delta_{\boldsymbol{\eta}^*}, x^\dagger) \leq C\delta.$$

*Proof.* By the minimizing property of $x^\delta_{\boldsymbol{\eta}^*}$, we have

$$t^*\psi_1(x^\delta_{\boldsymbol{\eta}^*}) + (1-t^*)\psi_2(x^\delta_{\boldsymbol{\eta}^*}) \leq t^*\psi_1(x^\dagger) + (1-t^*)\psi_2(x^\dagger).$$

Therefore, by the source condition, for some $\xi_{t^*} \in \partial(t^*\psi(x^\dagger) + (1 - t^*)\psi(x^\dagger))$ and $w_{t^*} \in Y$ such that $\xi_{t^*} = K^* w_{t^*}$, and the Cauchy-Schwarz inequality, we deduce

$$\begin{aligned} d_{\xi_{t^*}}(x^\delta_{\boldsymbol{\eta}^*}, x^\dagger) &= (t^*\psi_1(x^\delta_{\boldsymbol{\eta}^*}) + (1-t^*)\psi_2(x^\delta_{\boldsymbol{\eta}^*})) - (t^*\psi_1(x^\dagger) + (1-t^*)\psi_2(x^\dagger)) - \langle \xi_{t^*}, x^\delta_{\boldsymbol{\eta}^*} - x^\dagger \rangle \\ &\leq -\langle \xi_{t^*}, x^\delta_{\boldsymbol{\eta}^*} - x^\dagger \rangle = -\langle K^* w_{t^*}, x^\delta_{\boldsymbol{\eta}^*} - x^\dagger \rangle \\ &= -\langle w_{t^*}, K(x^\delta_{\boldsymbol{\eta}^*} - x^\dagger) \rangle \leq \|w_{t^*}\| \|K(x^\delta_{\boldsymbol{\eta}^*} - x^\dagger)\| \\ &\leq \|w_{t^*}\| \left( \|Kx^\delta_{\boldsymbol{\eta}^*} - y^\delta\| + \|y^\delta - Kx^\dagger\| \right) \leq (1 + c_m)\|w_{t^*}\|\delta. \end{aligned}$$

This shows the desired estimate. □

REMARK 3.2. *In the practical applications of the discrepancy principle, one needs to find the solution of a nonlinear equation in $\boldsymbol{\eta}$. The uniqueness of a solution to equation* (3) *is not guaranteed, and additional conditions need to be supplied for definiteness. Lastly, we would like to mention that the principle can be efficiently realized by the model function approach* [11].

**3.2. Balancing principle.** The discrepancy principle described earlier requires an estimate of the noise level $\delta$, which is not always available in practical applications. Therefore, it is of great interest to develop heuristic rules that do not require this knowledge. One such rule is the balancing principle, for which there are several variants, see [8] for details. The principle can be derived from the augmented Tikhonov (a-Tikhonov) regularization [10], which admits clear statistical interpretations as hierarchical Bayesian modeling. In particular, it provides the mechanism to automatically balance the penalty with the fidelity, see also Remark 3.4. The variant under consideration is due to [7], and has demonstrate very promising empirical results for several common single-parameter models [7]. Finally we remind the balancing principle discussed here should not be confused with the principle due to Lepskii which is sometimes also named balancing principle [15] and does require a precise knowledge of the noise level.

First we first sketch the a-Tikhonov regularization approach. For multi-parameter models, it can be derived analogously from Bayesian inference [10, 8], and the resulting a-Tikhonov functional $J(x, \tau, \{\lambda_i\})$ is given by

$$J(x, \tau, \{\lambda_i\}) = \tau \phi(x, y^\delta) + \boldsymbol{\lambda} \cdot \boldsymbol{\psi}(x) + \sum_i (\beta_i \lambda_i - \alpha_i \ln \lambda_i) + \beta_0 \tau - \alpha_0 \ln \tau,$$

which maximizes the posteriori probability density function $p(x, \tau, \{\lambda_i\}|y^\delta) \propto p(y^\delta|x, \tau, \{\lambda_i\}) \, p(x, \tau, \{\lambda_i\})$ under the assumption that the scalars $\{\lambda_i\}$ and $\tau$ have the Gamma distribution with known parameter pairs $(\alpha_i, \beta_i)$ and $(\alpha_0, \beta_0)$, respectively. Let $\eta_i = \frac{\lambda_i}{\tau}$. Then the necessary optimality condition of the a-Tikhonov functional is given by

$$\begin{cases} x_{\boldsymbol{\eta}}^\delta = \arg\min_x \left\{ \phi(x, y^\delta) + \boldsymbol{\eta} \cdot \boldsymbol{\psi}(x) \right\}, \\ \lambda_i = \dfrac{\alpha_i}{\psi_i(x_{\boldsymbol{\eta}}^\delta) + \beta_i}, \\ \tau = \dfrac{\alpha_0}{\phi(x_{\boldsymbol{\eta}}^\delta, y^\delta) + \beta_0}. \end{cases}$$

Upon assuming $\alpha_i = \alpha$ and $\beta_i = \beta$ for simplicity and letting $\gamma = \frac{\alpha_0}{\alpha}$, then we have the following system for $(x_{\boldsymbol{\eta}}^\delta, \boldsymbol{\eta})$

(5)
$$\begin{cases} x_{\boldsymbol{\eta}}^\delta = \arg\min_x \left\{ \phi(x, y^\delta) + \boldsymbol{\eta} \cdot \boldsymbol{\psi}(x) \right\}, \\ \eta_i = \dfrac{1}{\gamma} \dfrac{\phi(x_{\boldsymbol{\eta}}^\delta, y^\delta) + \beta}{\psi_i(x_{\boldsymbol{\eta}}^\delta) + \beta_0}. \end{cases}$$

Next, we give the promised balancing principle. The multi-parameter counterpart of the balancing principle given in [7] consists of minimizing

$$\Phi_\gamma(\boldsymbol{\eta}) = c_\gamma \frac{F^{2+\gamma}(\boldsymbol{\eta})}{\eta_1 \eta_2},$$

where the constant $c_\gamma = \frac{\gamma^\gamma}{(\gamma+2)^{\gamma+2}}$. We note that this constant $c_\gamma$ can be quite arbitrary, except for comparison with the criterion $\Psi_\gamma$ defined next. Another variant of the balancing principle reads

$$\Psi_\gamma(\boldsymbol{\eta}) = \phi(x_{\boldsymbol{\eta}}^\delta, y^\delta)^\gamma \psi_1(x_{\boldsymbol{\eta}}^\delta) \psi_2(x_{\boldsymbol{\eta}}^\delta),$$

which generalizes a criterion due to Reginska [6].

The relation between $\Phi_\gamma$ and $\Psi_\gamma$ is made explicit in the following result.

PROPOSITION 3.1. *Let the value function $F$ be twice continuously differentiable, $\partial_i F (i = 1, 2)$ do not vanish, and the Hessian $\nabla^2 F$ be nonsingular. Then the criteria $\Phi_\gamma$ and $\Psi_\gamma$ share the set of critical points, which are the solutions to the system*

$$(6) \qquad \gamma \eta_1 \psi_1(x_{\boldsymbol{\eta}}^\delta) = \gamma \eta_2 \psi_2(x_{\boldsymbol{\eta}}^\delta) = \phi(x_{\boldsymbol{\eta}}^\delta, y^\delta).$$

*Proof.* Setting the first-order derivatives of the criterion $\Phi_\gamma$ to zero gives

$$\nabla \Phi_\gamma(\eta) = c_\gamma \frac{F^{1+\gamma}(\boldsymbol{\eta})}{\eta_1 \eta_2} \left[ \begin{array}{c} (2+\gamma)\psi_1(x_{\boldsymbol{\eta}}^\delta) - \frac{F}{\eta_1} \\ (2+\gamma)\psi_2(x_{\boldsymbol{\eta}}^\delta) - \frac{F}{\eta_2} \end{array} \right] = \mathbf{0}.$$

This together with Lemma 2.3 gives $(2 + \gamma)\eta_i \psi_i(x_{\boldsymbol{\eta}}^\delta) = F$, $i = 1, 2$. Consequently, $\eta_1 \psi_1(x_{\boldsymbol{\eta}}^\delta) = \eta_2 \psi_2(x_{\boldsymbol{\eta}}^\delta)$, and thus system (6) holds. Meanwhile, by setting the first-order derivatives $\nabla \Psi_\gamma(\boldsymbol{\eta})$ of the the criterion $\Psi_\gamma$ to zero and noting Lemma 2.3, we get

$$\phi(x_{\boldsymbol{\eta}}^\delta, y^\delta)^{\gamma-1} \nabla^2 F \left[ \begin{array}{cc} \psi_2(x_{\boldsymbol{\eta}}^\delta) & \\ & \psi_1(x_{\boldsymbol{\eta}}^\delta) \end{array} \right] \left[ \begin{array}{c} -\gamma \eta_1 \psi_1(x_{\boldsymbol{\eta}}^\delta) + \phi(x_{\boldsymbol{\eta}}^\delta, y^\delta) \\ -\gamma \eta_2 \psi_2(x_{\boldsymbol{\eta}}^\delta) + \phi(x_{\boldsymbol{\eta}}^\delta, y^\delta) \end{array} \right] = \mathbf{0}.$$

By the assumption that the Hessian $\nabla^2 F$ is nonsingular and $\psi_i(x_{\boldsymbol{\eta}}^\delta)(i = 1, 2)$ do not vanish, we arrive at

$$\gamma \eta_i \psi_i(x_{\boldsymbol{\eta}}^\delta) - \phi(x_{\boldsymbol{\eta}}^\delta, y^\delta) = 0,$$

i.e., system (6). This concludes the proof. $\square$

REMARK 3.3. *Criterion $\Phi_\gamma$ makes only use of the value function $F(\eta)$, not of the derivatives of $F(\boldsymbol{\eta})$, which can be potentially multi-valued in case that the functional $J_{\boldsymbol{\eta}}$ has multiple minimizers. In contrast, the value function $F(\boldsymbol{\eta})$ is always continuous, see Lemma 2.1, and thus the optimization problem of minimizing $\Phi_\gamma$ over any bounded regions is always well-defined. For models with potentially nonunique minimizers, criterion $\Psi_\gamma$ and balancing principle, i.e., equation (6), are ill-defined, and the corresponding minimization formulations can be problematic. The criterion $\Phi_\gamma$ is advantageous then.*

REMARK 3.4. *Balancing principle is named after system (6): it attempts to balance the fidelity with the penalties with the parameter $\gamma$ being the relative weight. Comparing (6) with (5) shows clearly the intimate connections between the a-Tikhonov approach and the balancing principle: the a-Tikhonov approach builds in the principle automatically, and consequently the hierarchical Bayesian modeling is also balancing. Finally, we would like to remark that the balancing idea has been developed from other perspectives, see [8, Section 2.2] for details.*

The relation between the criteria $\Phi_\gamma$ and $\Psi_\gamma$ is made more precise in the following theorem: $\Psi_\gamma$ always lies below $\Phi_\gamma$, and thus at each local minimum, $\Phi_\gamma$ is sharper and numerically easier to locate.

THEOREM 3.6. *For any $\gamma > 0$, the following inequality holds*

$$\Psi_\gamma(\boldsymbol{\eta}) \le \Phi_\gamma(\boldsymbol{\eta}), \quad \forall \boldsymbol{\eta} > 0.$$

*The equality is achieved if and only if the balancing equation* (6) *is verified.*

*Proof.* Recall that for any $a, b, c \ge 0$ and $p, q, r > 1$ with $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$, there holds the generalized Young's inequality $abc \le \frac{a^p}{p} + \frac{b^q}{q} + \frac{c^r}{r}$, with equality holds if and only if $a^p = b^q = c^r$. Let $p = \frac{2+\gamma}{\gamma}$ and $q = r = 2 + \gamma$. Applying the inequality with $a = \phi^{\frac{\gamma}{2+\gamma}}(\eta_1\eta_2)^{-\frac{\gamma}{2(2+\gamma)}}$, $b = (\gamma\psi_1)^{\frac{1}{2+\gamma}}(\eta_1\eta_2^{-1})^{\frac{1}{2(2+\gamma)}}$ and $c = (\gamma\psi_2)^{\frac{1}{2+\gamma}}(\eta_1^{-1}\eta_2)^{\frac{1}{2(2+\gamma)}}$ gives

$$\phi^{\frac{\gamma}{2+\gamma}}\psi_1^{\frac{1}{2+\gamma}}\psi_2^{\frac{1}{2+\gamma}}(\eta_1\eta_2)^{-\frac{2}{2(2+\gamma)}}\gamma^{\frac{2}{2+\gamma}} \le \frac{\gamma}{2+\gamma}\frac{\phi + \boldsymbol{\eta}\cdot\boldsymbol{\psi}}{(\eta_1\eta_2)^{\frac{1}{2}}} = \frac{\gamma}{2+\gamma}\frac{F(\boldsymbol{\eta})}{(\eta_1\eta_2)^{\frac{1}{2}}}.$$

Hence

$$\phi^{\frac{\gamma}{2+\gamma}}\psi_1^{\frac{1}{2+\gamma}}\psi_2^{\frac{1}{2+\gamma}} \le \frac{\gamma^{\frac{\gamma}{\gamma+2}}}{2+\gamma}\frac{F(\boldsymbol{\eta})}{(\eta_1\eta_2)^{\frac{1}{2+\gamma}}}.$$

Therefore, we have

$$\Psi_\gamma(\boldsymbol{\eta}) \le \frac{\gamma^\gamma}{(2+\gamma)^{2+\gamma}}\frac{F^{2+\gamma}(\boldsymbol{\eta})}{\eta_1\eta_2} = \Phi_\gamma(\boldsymbol{\eta}).$$

The equality holds if and only if $a^p = b^q = c^r$, i.e.,

$$[\phi^{\frac{\gamma}{2+\gamma}}(\eta_1\eta_2)^{-\frac{\gamma}{2(2+\gamma)}}]^{\frac{2+\gamma}{\gamma}} = [(\gamma\psi_1)^{\frac{1}{2+\gamma}}(\eta_1\eta_2^{-1})^{\frac{1}{2(2+\gamma)}}]^{2+\gamma} = [(\gamma\psi_2)^{\frac{1}{2+\gamma}}(\eta_1^{-1}\eta_2)^{\frac{1}{2(2+\gamma)}}]^{2+\gamma}.$$

Simplifying this gives the balancing equation (6). This concludes the proof. $\square$

The following result shows an interesting property of a minimizer to Criterion $\Phi_\gamma$.

THEOREM 3.7. *At a local minimizer $\boldsymbol{\eta}^*$ to the function $\Phi_\gamma$, the partial derivatives of $F(\boldsymbol{\eta})$ exist.*

*Proof.* Assume that the assertion is not true, i.e., $\boldsymbol{\eta}^*$ is a discontinuity point of at least one $\psi_i$. Since $\boldsymbol{\eta}^*$ is a local minimizer, we have

$$\partial_i^- \Phi_\gamma(\boldsymbol{\eta}^*) \le 0 \quad \text{and} \quad \partial_i^+ \Phi_\gamma(\boldsymbol{\eta}^*) \ge 0.$$

In particular, this implies that $\partial_i^+ \Phi_\gamma(\boldsymbol{\eta}^*) - \partial_i^- \Phi_\gamma(\boldsymbol{\eta}^*) \ge 0$. Note that

$$\partial_i^+ \Phi_\gamma(\boldsymbol{\eta}^*) - \partial_i^- \Phi_\gamma(\boldsymbol{\eta}^*) = (2+\gamma)c_\gamma \frac{1}{\eta_1^*\eta_2^*} F^\gamma(\boldsymbol{\eta}^*)\left[\partial_i^+ F(\boldsymbol{\eta}^*) - \partial_i^- F(\boldsymbol{\eta}^*)\right]$$

and consequently $\partial_i^+ F(\boldsymbol{\eta}^*) - \partial_i^- F(\boldsymbol{\eta}^*) \ge 0$. This is in contradiction with the fact that at a discontinuity point $\boldsymbol{\eta}^*$, $\partial_i^+ F(\boldsymbol{\eta}^*) - \partial_i^- F(\boldsymbol{\eta}^*) < 0$ by the monotonicity of the function $\psi_i(x_{\boldsymbol{\eta}}^\delta)$ with respect to $\eta_i$. $\square$

---

**Algorithm 1** Fixed point algorithm I.

---

1: Choose $\gamma$, $\boldsymbol{\eta}^0$ and set $k = 0$.

2: **repeat**

3:     Solve for $x^{k+1}$ by the Tikhonov regularization method

$$x^{k+1} = \arg\min_x \left\{ \phi(x, y^\delta) + \boldsymbol{\eta}^k \cdot \boldsymbol{\psi}(x) \right\}.$$

4:     Update the regularization parameter $\boldsymbol{\eta}^{k+1}$ by

$$\eta_1^{k+1} = \frac{1}{1+\gamma} \frac{\phi(x^{k+1}, y^\delta) + \eta_2^k \psi_2(x^{k+1})}{\psi_1(x^{k+1})},$$

$$\eta_1^{k+1} = \frac{1}{1+\gamma} \frac{\phi(x^{k+1}, y^\delta) + \eta_1^k \psi_1(x^{k+1})}{\psi_2(x^{k+1})}.$$

5: **until** A stopping criterion is satisfied.

---

Now we present an a posteriori error estimate for Criterion $\Phi_\gamma$ when $Y$ is a Hilbert space and $\phi(x, y^\delta) = \|Kx - y^\delta\|^2$ and convex penalties. The proof will be presented elsewhere, and we also refer to [7]. Theorem 3.8 provides one a posteriori way to check the automatically determined (vector-valued) regularization parameter, and partially justifies the criterion theoretically.

THEOREM 3.8. *Let the following source condition be satisfied for the exact solution* $x^\dagger$: *for any* $t \in [0,1]$ *there exists a* $w_t \in Y$

$$\xi_t \in \partial \left( t\psi_1(x^\dagger) + (1-t)\psi_2(x^\dagger) \right) \quad and \quad \xi_t = K^* w_t.$$

*Then for every* $\boldsymbol{\eta}^*$ *determined by the criterion* $\Phi_\gamma$, *there exists some constant* $C$ *such that*

$$d_{\xi_{t^*}}(x_{\boldsymbol{\eta}^*}^\delta, x^\dagger) \le C \left( \|w_{t^*}\| + \frac{F^{2+\gamma}(\delta\boldsymbol{e})}{F^{2+\gamma}(\boldsymbol{\eta}^*)} \right) \max(\delta, \delta_*),$$

*where* $\boldsymbol{e} = (1,1)^{\mathrm{T}}$, $\delta_* = \|Kx_{\boldsymbol{\eta}^*}^\delta - y^\delta\|$, *and* $t^* = \eta_1^*/(\eta_1^* + \eta_2^*)$.

Finally, we present two algorithms, see Algorithms 1 and 2, for computing a minimizer of Criterion $\Phi_\gamma$. The algorithms are of fixed point type, and can be regarded as natural extensions of the fixed point algorithm in [7]. Practically, the algorithms merit a very steady and fast convergence.

**4. Numerical experiments.** This part presents numerical results for three examples, which are integral equations of the first kind with kernel $k(s,t)$ and solution $x^\dagger(t)$, to illustrate features of multi-parameter models. The discretized linear system takes the form $\mathbf{K}\mathbf{x}^\dagger = \mathbf{y}^\dagger$. The data $\mathbf{y}^\dagger$ is corrupted by noises, i.e., $y_i^\delta = y_i^\dagger + \max_i\{|y_i^\dagger|\}\varepsilon\xi_i$, where $\xi_i$ are standard Gaussian variables and $\varepsilon$ refers to the relative noise level. The fidelity $\phi$ is taken to be the standard least-squares fitting. We present only the numerical results for Algorithm II, as Algorithm I exhibits similar convergence behavior. The initial guess is always taken to be $1 \times 10^{-3}$, and it is stopped if the relative change of $\boldsymbol{\eta}$ is smaller than $1.0 \times 10^{-3}$. The parameter $\gamma$ in Criterion $\Phi_\gamma$ is determined by a two-step procedure [7]: The initial guess for $\gamma$ is set to 5, and then it is automatically adjusted according to the estimate noise level.

---

**Algorithm 2** Fixed point algorithm II.

---

1: Choose $\gamma$, $\boldsymbol{\eta}^0$ and set $k = 0$.

2: **repeat**

3:    Solve for $x^{k+1}$ by the Tikhonov regularization method

$$x^{k+1} = \arg\min_x \left\{ \phi(x, y^\delta) + \boldsymbol{\eta}^k \cdot \boldsymbol{\psi}(x) \right\}.$$

4:    Update the regularization parameter $\boldsymbol{\eta}^{k+1}$ by

$$\eta_1^{k+1} = \frac{1}{\gamma} \frac{\phi(x^{k+1}, y^\delta)}{\psi_1(x^{k+1})},$$

$$\eta_2^{k+1} = \frac{1}{\gamma} \frac{\phi(x^{k+1}, y^\delta)}{\psi_2(x^{k+1})}.$$

5: **until** A stopping criterion is satisfied.

---

TABLE 1
*Numerical results for Example 1.*

| $\epsilon$ | $\boldsymbol{\eta}_{\text{b}}$ | $\boldsymbol{\eta}_{\text{o}}$ | $\eta_{\text{h1}}$ | $\eta_{\text{tv}}$ | $e_{\text{b}}$ | $e_{\text{o}}$ | $e_{\text{h1}}$ | $e_{\text{tv}}$ |
|---|---|---|---|---|---|---|---|---|
| 5e-2 | (3.44e-3,5.75e-3) | (2.36e-4,2.14e-3) | 5.68e-4 | 9.27e-3 | 3.31e-2 | 2.66e-2 | 3.97e-2 | 1.07e-1 |
| 5e-3 | (1.03e-4,1.83e-4) | (2.19e-5,3.70e-4) | 6.81e-5 | 4.85e-4 | 2.27e-2 | 1.10e-2 | 2.69e-2 | 9.48e-2 |
| 5e-4 | (3.32e-6,6.12e-6) | (2.89e-6,5.07e-5) | 1.26e-6 | 6.08e-5 | 1.25e-2 | 8.85e-3 | 1.38e-2 | 4.48e-2 |
| 5e-5 | (1.07e-7,2.04e-7) | (7.04e-8,5.23e-6) | 1.14e-7 | 4.06e-6 | 6.82e-3 | 5.53e-3 | 9.40e-3 | 1.68e-2 |
| 5e-6 | (3.01e-9,5.77e-9) | (2.06e-10,6.65e-9) | 6.01e-10 | 2.24e-7 | 4.50e-3 | 2.89e-3 | 5.28e-3 | 5.12e-3 |

**4.1. $H^1$-TV model.** EXAMPLE 1. *Let $\zeta(t) = \chi_{|t| \le 3}(1 + \cos\frac{\pi t}{3})$, and the kernel $k$ is given by $k(s,t) = \zeta(s - t)$. The exact solution $x^\dagger$ is shown in Fig. 1, and the integration interval is $[-6, 6]$. The solution $x^\dagger$ exhibits both flat and smoothly varying regions, and thus we adopt two penalties $\psi_1(x) = \frac{1}{2}|x|_{H^1}^2$ and $\psi_2(x) = |x|_{TV}$ for preserving their distinct features. The size of the problem is 100.*

The numerical results are summarized in Table 1. In the table, the subscripts b and o refer to the balancing principle and the optimal choice, i.e., the value giving the smallest reconstruction error, respectively. The results for single-parameter models are indicated by subscripts h1 and tv, and the respective penalty parameter shown in Table 1 is the optimal one. The accuracy of the results is measured by the relative $L^2$ error $e = \|x - x^\dagger\|/\|x^\dagger\|$. A first observation is that the error $e_{\text{b}}$, by the balancing principle for the proposed model $H^1$-$TV$ is smaller than the optimal choice for either $H^1$ or $TV$ penalty. This illustrates clearly the benefit of using multi-parameter model. Interestingly, the balancing principle gives an error fairly close to the optimal one, and the error decreases as the noise level decreases.

The numerical results for Example 1 with $\varepsilon = 5\%$ noise is shown in Fig. 1. In particular, the classical $H^1$ smoothness penalty fails to restore the flat region satisfactorily, whereas the $TV$ approach suffers from stair-case effect in the gray region and reduced magnitude in the flat region, see Fig. 1. In contrast, the proposed $H^1$-$TV$ model can preserve the magnitude of flat region while reconstruct the gray region excellently. Therefore, it indeed combines the strengths of both $H^1$ and $TV$ models, and is suitable for restoring images with both flat and gray regions. The criterion $\Phi_\gamma$ is numerically well-behaved: there is a distinct local minimum, and it is numerically easy to minimize, see Fig. 2. Finally, we would like to remark that the algorithm converge rapidly with the convergence achieved in five iterations, see Fig. 2.
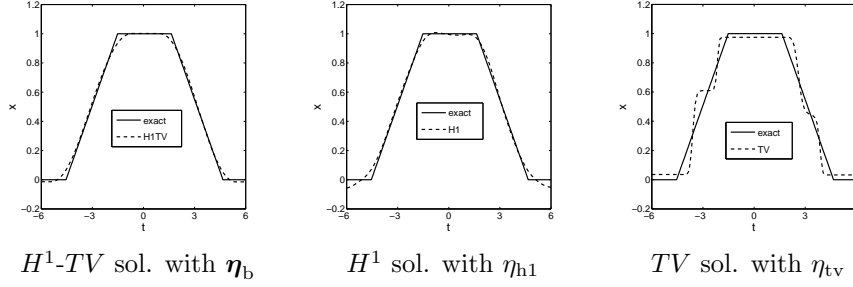
$H^1$-$TV$ sol. with $\boldsymbol{\eta}_{\mathrm{b}}$     $H^1$ sol. with $\eta_{\mathrm{h1}}$     $TV$ sol. with $\eta_{\mathrm{tv}}$

Fig. 1. *Numerical results for Example 1 with 5% noise.*



$\Phi_\gamma(\boldsymbol{\eta})$       convergence of Algorithm II
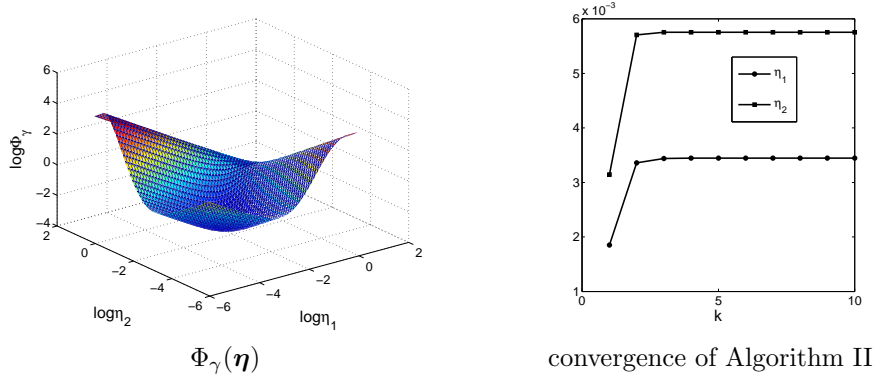
Fig. 2. *Numerical results for Example 1 with 5% noise.*

**4.2. $\ell^1$-$\ell^2$ model.** EXAMPLE 2. *The kernel $k$ is given by $k(s,t) = \frac{1}{4}\left(\frac{1}{16} + (s-t)^2\right)^{-\frac{3}{2}}$, the exact solution $x^\dagger$ consists of two bumps and it is shown in Fig. 3. The penalties are $\psi_1(x) = \|x\|_{\ell^1}$ and $\psi_2(x) = \frac{1}{2}\|x\|_{\ell^2}^2$ to retrieve the groupwise sparsity structure. The integration interval is $[0,1]$. The size of the problem is 100.*

The numerical results for this example are show in Table 2 and Fig. 3. Here we are interested in the group structure of the solution with minimal number of influencing factors (nonzero coefficients). Again, we observe that the elastic-net compares favorably with the conventional $\ell^1$ and $\ell^2$ penalties in terms of the error, and the balancing principle can give reasonable estimate for the optimal choice. The conventional $\ell^2$ solution contains almost no zero entries, and thus it fails to distinguish between influencing and noninfluencing coefficients, i.e., identifying relevant factors. This difficulty is partially remedied by the $\ell^1$ model in that many entries of the $\ell^1$ solution are zero. Therefore, some relevant factors are correctly identified. However, it tends to select only some instead of all relevant factors, i.e., group structure. The elastic-net model combines the best of both $\ell^1$ and $\ell^2$ models, and it achieves the desired goal of identifying the group structure. Moreover, the magnitude assigned to the coefficients are reasonable compared to others. The algorithm converges quickly within five iterations.

**4.3. 2D image deblurring.** EXAMPLE 3. *The penalties are $\psi_1(x) = \|x\|_{\ell^1}$ and $\psi_2(x) = \frac{1}{2}\|x\|_{\ell^2}^2$. The kernel $k$ performs standard Gaussian blur with standard deviation 1 and blurring width 5. The exact solution $x^\dagger$ is shown in Fig. 5. The size*
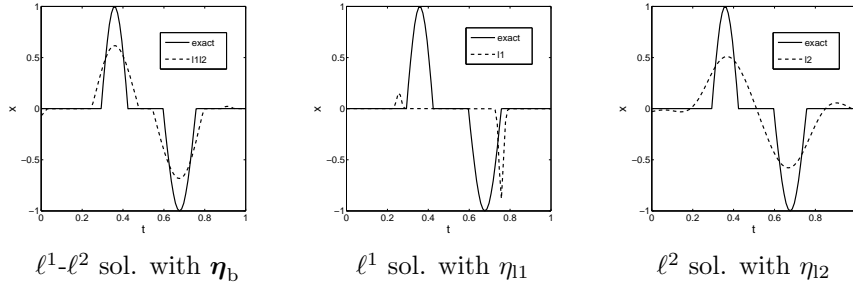
$\ell^1$-$\ell^2$ sol. with $\boldsymbol{\eta}_{\mathrm{b}}$       $\ell^1$ sol. with $\eta_{l1}$       $\ell^2$ sol. with $\eta_{l2}$

FIG. 3. *Numerical results for Example 2 with 5% noise.*



$\Phi_\gamma(\boldsymbol{\eta})$       convergence of Algorithm II
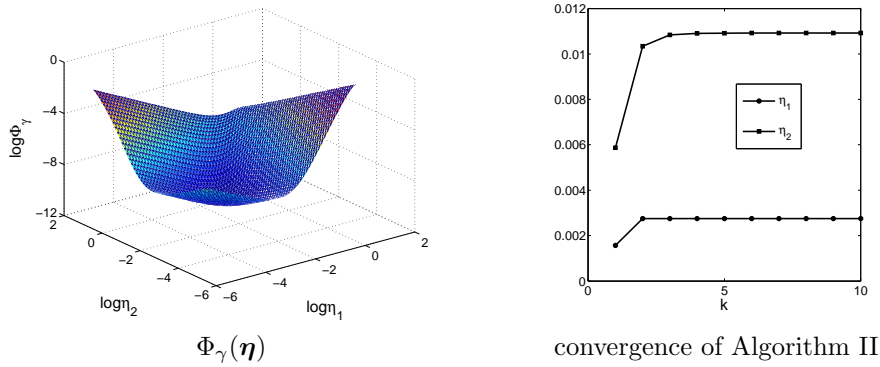
FIG. 4. *Numerical results for Example 2 with 5% noise.*

*of the image is* $50 \times 50$.

This example showcases a more realistic problem of image deblurring. Here one half of the data points are retained. The reconstructions for 1% noise are shown in Fig. 5. The $\ell^1$ solution is more spiky, and neighboring pixels more or less act independently. In particular, due to missing data, there are some missing pixels in the blocks and the cross to be recovered. In contrast, the $\ell^2$ solution is more blockwise, but there are many nonzero coefficients indicated by the small spurious oscillations in the background. The elastic-net model achieves the best of the two: retaining the block structure with only fewer spurious nonzero coefficients. This is deemed important in medical imaging, e.g., classification. The numbers are more telling: $e_{\mathrm{b}} = 2.99 \times 10^{-1}$, $e_{\mathrm{o}} = 2.44 \times 10^{-1}$, $e_{l1} = 9.21 \times 10^{-1}$, and $e_{l2} = 3.42 \times 10^{-1}$. Therefore, the error $e_{\mathrm{b}}$ for elastic-net agrees well with the optimal choice, and it is smaller than that with the optimal choices for both $\ell^1$ and $\ell^2$ models.

**5. Concluding remarks.** We have studied theoretical properties of multi-parameter Tikhonov regularization. Some properties, e.g., monotonicity, concavity, asymptotic and differentiability, of the value function, were established. The discrepancy principle is partially justified in terms of consistency and convergence rates, however, the regularization parameter is not uniquely determined, which partially limits its practical application. It is of interest to develop auxiliary rules, which is currently under investigation. In contrast, the balancing principle allows justifications in terms of a posteriori error estimate and efficient numerical implementation. The numerical experiments show that multi-parameter models can significantly improve

TABLE 2
*Numerical results for Example 2.*

| $\epsilon$ | $\boldsymbol{\eta}_{\mathrm{b}}$ | $\boldsymbol{\eta}_{\mathrm{o}}$ | $\eta_{\mathrm{l1}}$ | $\eta_{\mathrm{l2}}$ | $e_{\mathrm{b}}$ | $e_{\mathrm{o}}$ | $e_{\mathrm{l1}}$ | $e_{\mathrm{l2}}$ |
|---|---|---|---|---|---|---|---|---|
| 5e-2 | (2.75e-3,1.09e-2) | (3.16e-3,1.32e-3) | 2.96e0 | 3.34e-3 | 4.18e-1 | 8.72e-2 | 1.04e0 | 4.59e-1 |
| 5e-3 | (9.16e-5,2.86e-4) | (2.46e-4,1.07e-4) | 1.03e-4 | 3.06e-5 | 2.09e-1 | 1.24e-2 | 8.97e-1 | 2.90e-1 |
| 5e-4 | (2.82e-6,7.48e-6) | (2.34e-5,1.14e-5) | 1.30e-5 | 4.08e-6 | 5.76e-2 | 7.98e-3 | 6.18e-1 | 2.17e-1 |
| 5e-5 | (8.89e-8,2.26e-7) | (2.27e-6,1.06e-6) | 1.24e-6 | 3.84e-8 | 1.57e-2 | 4.71e-3 | 4.85e-1 | 1.66e-1 |
| 5e-6 | (2.79e-9,7.07e-9) | (1.66e-7,1.03e-7) | 4.12e-9 | 1.41e-9 | 1.27e-2 | 2.27e-3 | 2.61e-1 | 9.55e-2 |



exact $x^{\dagger}$          $\ell^1$-$\ell^2$ sol. with $\boldsymbol{\eta}_{\mathrm{o}}$ =(1.25e-2,1.29e-3)

$\ell^1$-$\ell^2$ sol. with $\boldsymbol{\eta}_{\mathrm{b}}$ =(1.14e-3, 1.12e-3)          $\ell^1$ sol. with $\eta_{\mathrm{l1}}$=5.30e-1          $\ell^2$ sol. with $\eta_{\mathrm{l2}}$ =3.31e-3
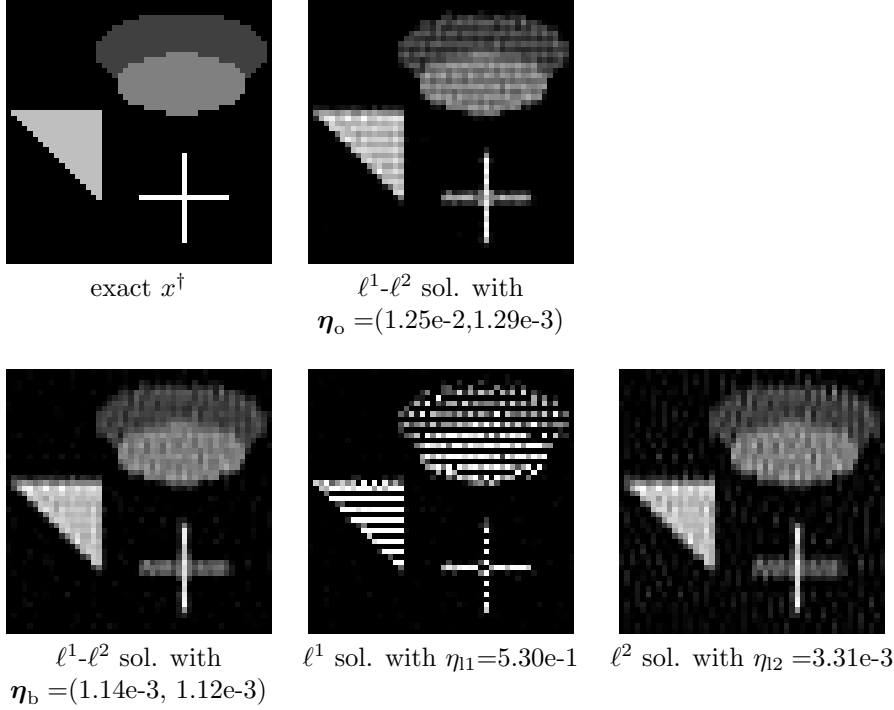
FIG. 5. *Numerical results for Example 3 with* 1% *noise.*

the reconstruction quality and the balancing principle can give reasonable results in comparison with the optimal choice in a computationally efficient way. The two proposed algorithms for computing the parameters of the balancing principle deliver excellent convergence behavior. However, a rigorous convergence analysis remains to be established.

REFERENCES

[1] M. BELGE, M. E. KILMER, AND E. L. MILLER, *Efficient determination of multiple regularization parameters in a generalized L-curve framework*, Inverse Problems, 18 (2002), pp. 1161–1183.
[2] C. BREZINSKI, M. REDIVO-ZAGLIA, G. RODRIGUEZ, AND S. SEATZU, *Multi-parameter regularization techniques for ill-conditioned linear systems*, Numer. Math., 94 (2003), pp. 203–228.

[3]  D. H. Brooks, G. F. Ahmad, R. S. Macleod, and G. M. Maratos, *Inverse electrocardiography by simultaneous imposition of multiple constraints*, IEEE Trans. Biomed. Eng., 46 (1999), pp. 3–18.

[4]  Z. Chen, Y. Lu, Y. Xu, and H. Yang, *Multi-parameter Tikhonov regularization for linear ill-posed operator equations*, J. Comput. Math., 26 (2008), pp. 37–55.

[5]  C. De Mol, E. De Vito, and L. Rosasco, *Elastic-net regularization in learning theory*, J. Complexity, 25 (2009), pp. 201–230.

[6]  H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 1996.

[7]  K. Ito, B. Jin, and T. Takeuchi, *A regularization parameter for nonsmooth Tikhonov regularization*, Tech. Report UTMS 2010-3, Graduate School of Mathematical Sciences, University of Tokyo, 2010.

[8]  K. Ito, B. Jin, and J. Zou, *A new choice rule for regularization parameters in Tikhonov regularization*, Appl. Anal., (2010), p. in press.

[9]  B. Jin, D. A. Lorenz, and S. Schiffler, *Elastic-net regularization: error estimates and active set methods*, Inverse Problems, 25 (2009), p. 115022 (26pp).

[10] B. Jin and J. Zou, *Augmented Tikhonov regularization*, Inverse Problems, 25 (2009), pp. 025001, 25.

[11] ———, *Iterative parameter choice by discrepancy principle*, Tech. Report 2010-02(369), Department of Mathematics, Chinese University of Hong Kong, 2010.

[12] S. Lu and S. V. Pereverzev, *Multi-parameter regularization and its numerical regularization*. Numer. Math., doi: 10.1007/s00211-010-0318-3.

[13] S. Lu, S. V. Pereverzev, Y. Shao, and U. Tautenhahn, *Discrepancy curves for multi-parameter regularization*, J. Inv. Ill-Posed Probl., 18 (2010), pp. 655–676.

[14] Y. Lu, L. Shen, and Y. Xu, *Multi-parameter regularization methods for high-resolution image reconstruction with displacement errors*, IEEE Trans. Circuits Syst. I. Regul. Pap., 54 (2007), pp. 1788–1799.

[15] P. Mathé, *The Lepskii principle revisited*, Inverse Problems, 22 (2006), pp. L11–L15.

[16] V. A. Morozov, *On the solution of functional equations by the method of regularization*, Sov. Math. Dokl., 7 (1966), pp. 414–417.

[17] I. M. Stephanakis, *Regularized image restoration in multiresolution spaces*, Opt. Eng., 36 (1997), pp. 1738–1744.

[18] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*, John Wiley, New York, 1977.

[19] P. Xu, Y. Fukuda, and Y. Liu, *Multiple parameter regularization: numerical solutions and applications to the determination of geopotential from precise satellite orbits*, J. Geod., 80 (2006), pp. 17–27.

[20] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B, 67 (2005), pp. 301–320.