Dipartimento di Informatica e
Scienze dell'Informazione

Minimization of Tikhonov Functional: the Continuous Setting

Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto,
Michele Piana, Alessandro Verri

Technical Report **DISI-TR-03-14**

DISI, Università di Genova
v. Dodecaneso 35, 16146 Genova, Italy http://www.disi.unige.it/

Abstract

In these notes we study the explicit form of the minimizer of the functional obtained adding a penalty term to the expected risk. Our study provides a quantitative version of the well known representer theorem and hold both for regression and classification under very mild and natural assumption.

1. Definitions

We assume that the pair $(\mathbf{x}, y)$ is in $Z = X \times Y$, where X is a compact subset of $\mathbb{R}^d$ and Y is a compact subset of $\mathbb{R}$ (for regression $Y = [a, b]$, for binary classification $Y = \{-1, 1\}$). We assume that Z is endowed with a probability measure ρ that can be decomposed in the following way $\rho(y, \mathbf{x}) = \nu(\mathbf{x})\rho_{\mathbf{x}}(y)$, where $\nu(\mathbf{x})$ is the marginal probability measure on X and $\rho_{\mathbf{x}}(y)$ the conditional probability on Y.

Let the hypothesis space $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS) with a continuous kernel $K : X \times X \to \mathbb{R}$ (Aronszajn, 1950). We recall that $\mathcal{H}$ is defined as the unique Hilbert space of continuous functions on X such that

$$f(\mathbf{x}) = \langle f, K_{\mathbf{x}} \rangle_{\mathcal{H}}, \tag{1}$$

where, for all $\mathbf{x} \in X$, $K_{\mathbf{x}}$ is the function on X defined by $K_{\mathbf{x}}(\mathbf{s}) = K(\mathbf{x}, \mathbf{s})$.

The loss function $V(y, f(\mathbf{x}))$ is the price we are willing to pay by using $f(\mathbf{x})$ to predict the correct label y (see the following for the mathematical properties we assume on V).

Given a function f its expected risk is defined as

$$I[f] = \int_{X \times Y} V(y, f(\mathbf{x})) d\rho(y, \mathbf{x}).$$

In the following we address the problem of finding the explicit form of the solution of the problem

$$\min_{f \in \mathcal{H}} \{I[f] + \lambda \|f\|^2_{\mathcal{H}}\}, \tag{2}$$

The main result of our study is expressed in the following theorem

Theorem 1 *Given $\lambda > 0$ the problem*

$$\inf_{f \in \mathcal{H}} \{ \int_{X \times Y} V(y, f(\mathbf{x})) d\rho(y, \mathbf{x}) + \lambda \|f\|^2_{\mathcal{H}} \}$$

admits a unique solution $f_\lambda \in \mathcal{H}$ *given by*

$$f_\lambda = -\frac{1}{2\lambda}\int_{X\times Y} K(\mathbf{s},\mathbf{x})\alpha(y,\mathbf{x})d\rho(y,\mathbf{x})$$

where

$$\alpha(y,\mathbf{x}) \in (\partial V)(f_\lambda(\mathbf{x}), y)$$

here (∂V) *is the subgradient of* V *with respect to its second argument.*

The plan of these notes is as follows. In section **??** we discuss what has already been done on this subject and the advantages of our work. In section 2 we discuss the hypothesis that we assume on the loss V. In section 3 we recall some notions from convex analysis in infinite dimensional spaces. In section 4 we sketch informally the proof of the main theorem and in section 5 we give the mathematical proof. Finally in section **??** we make some comments on the obtainded results.

2. Hypothesis

The main mathematical properties of the functional in (2) are direct consequences of the assumption that we make on the loss function V. In the following we state and briefly discuss the latter.

Definition 2 *A loss function* V *is a function* $V : Y \times \mathbb{R} \to [0,\infty[$*, s.t.*

1. V *is a measurable function*
2. $\forall\, y \in Y$, $V(y,\cdot)$*, is a convex function on* $\mathbb{R}$
3. $\exists\, a, b \in\ ]0,+\infty[$ *s.t.*
$$|V(y,w)| \le a|w|^2 + b \qquad (3)$$
4. *the dependence of* $V(y,w)$ *on its arguments is either of the form* $V(yw)$ *or* $V(w-y)$*.*

Remark 3 *In condition* 3) *we assume a square power dependency but with little work the following results can be obtained for an arbitrary* p *power dependency (see (Ekeland and Turnbull, 1983)).*

Let us discuss the last three conditions. The convexity hypothesis on V is very natural. In fact if condition 2) holds it is straightforward to check that the functional in (2) is strictly convex and this turns out to be fundamental to ensure existence and uniqueness of the minimizer. On the other hand, it is not a restrictive assumption indeed, since it is satisfied by all the loss functions commonly used.

Condition 3) is a more technical hypothesis we need in order to prove the continuity of the functional in (2). In fact we recall that in finite dimensional spaces convexity implies continuity, but this is no longer true as we pass to infinite dimensional spaces. In this case a possible way to ensure the continuity is to require some Lipschitz-like condition on V. It is well known that if V is convex it is locally Lipschitz but in order to ensure continuity we need some stronger condition. A possibility is to require V to be globally Lipschitz (see (Steinwart, 2002)), but this would cut off some very important loss such as the square one. Condition 3) provides a weaker and hence more general condition and has an intuitive interpretation. If condition 3) holds V is locally Lipschitz and the Lipschitz constant grows at most linearly with the considered interval. This means that we admit loss functions that have at most a square power dependency on their first argument. Again this allows us to handle all the commonly used loss functions with the only exception of the exponential loss function $V(y, f(\mathbf{x})) = e^{-yf(\mathbf{x})}$.

Condition 4) will be useful in the proof of Lemma 10. It is fulfilled by all the loss functions used in practice, in particular the dependence $V(yw)$ refers to the case of classification, while $V(w-y)$ to regression.

3. Convex functions in infinite dimensional spaces

In this section we recall some results of convex analysis that we will use in the following. First we give some basic definitions and results. Second we introduce some notions of convex analysis for integral functionals that will be fundemental to evaluate the subgradient of the expected risk $I[f]$.

3.1 Basic definitions

We briefly recall some properties of convex functions defined on a Hilbert space $\mathcal{H}$. For a detailed review see, for example (Ekeland and Turnbull, 1983).

A function $F : \mathcal{H} \to \mathbb{R}$ is *convex* if

$$F(tv + (1-t)w) \leq tF(v) + (1-t)F(w),$$

for all $v, w \in \mathcal{H}$ and $t \in [0,1]$ (if the strict inequality holds for $t \in (0,1)$, F is called strictly convex).

The *subgradient* of F at point $v_0 \in \mathcal{H}$ is the subset of $\mathcal{H}$ given by

$$(\partial F)_{v_0} = \{w \in \mathcal{H} \,|\, F(v) \geq F(v_0) + \langle w, v - v_0 \rangle_{\mathcal{H}} \ \ \forall v \in \mathcal{H}\}. \quad (4)$$

If F is differentiable in v_0, the subgradient reduces to the usual gradient $F'(v_0)$ (Ekeland and Turnbull, 1983, Prop. $III.2.8$) and inequality (4) is the usual definition of convex function

$$F(v) \geq F(v_0) + \langle F'(v_0), v - v_0 \rangle .$$

If $\mathcal{H} = \mathbb{R}^2$, inequality (4) has a simple geometrical interpretation. A vector (w_1, w_2) is in the subgradient if and only if the plane

$$z = F(x_0, y_0) + w_1(x - x_0) + w_2(y - y_0)$$

is under the graph $z = F(x, y)$. In the following we will make large use of the next result.

Remark 4 *If F is a convex function on $\mathbb{R}$, then it is continuous (Ekeland and Turnbull, 1983, Cor. $III.1.2$), left and right derivatives always exist with $F'_-(x) \leq F'_+(x)$ (Ekeland and Turnbull, 1983, Prop. $III.2.7$), and $(\partial F)_x = [F'_-(x), F'_+(x)]$.*

We need the following facts extending the linearity, extremality condition and chain rule properties for the gradient of differentiable functions to the subgradient of convex ones.

Proposition 5 *Let F, F_1 and F_2 be convex functions then following facts hold:*

a) let F_1 and F_2 be continuous convex functions on $\mathcal{H}$ and $a, b \geq 0$, then $F = aF_1 + bF_2$ is convex and

$$(\partial F)_v = a(\partial F_1)_v + b(\partial F_2)_v;$$

b) F has a minimum point at v if and only if $0 \in (\partial F)_v$;

c) if F is defined on $\mathbb{R}$ and $w \in \mathcal{H}$, then the function on $\mathcal{H}$

$$v \to F(\langle v, w \rangle)$$

is convex, continuous and its subgradient at v_0 is given by

$$[F'_-(\langle v_0, w \rangle), F'_+(\langle v_0, w \rangle)] \; w.$$

Proof See Prop. $III.2.9$ and Cor. $III.2.1$ in Ekeland and Turnbull (1983) for item a), Prop. $III.3.1$ in Ekeland and Turnbull (1983) for item b) and Prop. $III.2.12$ in Ekeland and Turnbull (1983) for item c). ■

Remark 6 *If F is a convex continuous function such that*

$$\lim_{\|v\|_{\mathcal{H}}\to\infty} F(v) = +\infty.$$

then F has a minimizer Ekeland and Turnbull (1983, Prop. $II.4.6$). If F is strictly convex, the minimizer is unique.

3.2 Convex analysis for integral functionals

The defintions and results of this section are all taken from (Ekeland and Turnbull (1983), Section $III.5$) with some minor modifications. Let $W : \mathbb{R}\times X \to [0, +\infty[\cup \{+\infty\}$ be convex and lower semi-continuous in its first argument, we define the functional $I_0 : L^2(X, \nu) \to [0, +\infty[\ \cup \ \{+\infty\}$ by

$$I_0[f] = \int_X W(f(\mathbf{x}), \mathbf{x}) d\nu(\mathbf{x}). \tag{5}$$

The above functional is known as the *Nemitski* functional associated to W. It is easy to prove that I_0 is convex since $W(\cdot, \mathbf{x})$ is convex ((Ekeland and Turnbull, 1983) refer to Theorem $II.5.1$). The following proposition discusses the continuity of I_0.

Proposition 7 *If $W(\cdot, \mathbf{x})$ is continuous for all $\mathbf{x} \in X$ and $|W(w, \mathbf{x})| \le a|w|^2 + b(\mathbf{x})$ for all $\mathbf{x} \in X$ and $w \in \mathbb{R}$, where $a \in [0, \infty[$ and $b \in L^1(X, \nu)$, then I_0 is continuous.*

For a proof see (Ekeland and Turnbull (1983), Prop. $III.5.1$).
Finally next proposition provides us with a straightforward method to study the subgradient (∂I_0).

Proposition 8 *If there is an element $\hat{f} \in L^\infty(X, \nu)$ s.t. $I_0[\hat{f}] < \infty$ then for all $f \in L^2(X, \nu)$*

$$(\partial I_0)(f) = \{f^* \in L^2(X, \nu) : \ f^*(\mathbf{x}) \in (\partial W)(f(\mathbf{x}), \mathbf{x}) \ a.e. \ [\nu]\}. \tag{6}$$

For a proof see (Ekeland and Turnbull (1983), Prop. $III.5.3$).

Remark 9 *Since in our setting X is compact, it follows that in Proposition 7 a constant is a suitable function $b(\mathbf{x})$ in $L^1(X, \nu)$. A condition of this form will be obtained from our general hypothesis in Eq. (3).*

4. Informal discussion

We now try to sketch the idea behind our main theorem and the main steps in its proof. First of all we recall that the functional that we have to minimize is defined by

$$J[f] = \int_{X\times Y} V(y, f(\mathbf{x}))d\rho(y,\mathbf{x}) + \lambda \|f\|_{\mathcal{H}}^2 . \tag{7}$$

The idea in order to calculate the minimizer of the above functional is to evaluate its subgradient $(\partial J)(f)$ and then set $(\partial J)(f)$ equal to zero (see Section 3.1). The main problem is to evaluate the subgradient of the first term

$$I[f] = \int_{X\times Y} V(y, f(\mathbf{x}))d\rho(y,\mathbf{x})$$

To evaluate the subgradient $(\partial J)(f)$ we would like to use the result about the subgradient of the Nemitski functional (see Section 3.2). Roughly speaking, to do this we have to get rid of the dependence on y and modify the results of Section 3.2 since the functional we are dealing with is not defined on $L^2(X,\nu)$ but on $\mathcal{H}$. Now the idea is to introduce

$$W(f(\mathbf{x}),\mathbf{x}) = \int_Y V(y, f(\mathbf{x}))d\rho_{\mathbf{x}}(y) \tag{8}$$

so that we can rewrite I as

$$I[f] = \int_{X\times Y} V(y, f(\mathbf{x}))d\rho(y,\mathbf{x}) = \int_X W(f(\mathbf{x}),\mathbf{x})d\nu(\mathbf{x})$$

where we used (8) and the fact that the probability measure ρ decomposes in $\rho(y,\mathbf{x}) = \rho_{\mathbf{x}}(y)\nu(\mathbf{x})$. The above functional has exactly the form of the Nemitski functional and we only have to cope with the passage from $L^2(X,\nu)$ to $\mathcal{H}$, that is we have to pass from the functional I_0 to I. Summing up the main steps of the proof are the following. In Lemma 10 we show that the assumptions we made on the loss function V imply that W satisfies the condition of Section 3.2. In Lemma 11 we prove the continuity of I_0 and evaluate its subgradient. Finally in Theorem 12 we pass from $L^2(X,\nu)$ to $\mathcal{H}$ and evaluate the explicit form of the minimizer of (7).

5. Proof

Next lemma studies the main properties of the function W.

Lemma 10 *The function* $W : \mathbb{R} \times X \to [0, +\infty[\ \cup\ \{+\infty\}$ *defined by*

$$W(f(\mathbf{x}), \mathbf{x}) = \int_Y V(y, f(\mathbf{x})) d\rho_{\mathbf{x}}(y)$$

has the following properties:

1. $\forall\ \mathbf{x} \in X$, $W(\cdot, \mathbf{x})$ *is convex.*

2. W *is a measurable function.*

3. W *is finite. Furthermore let* $a, b \in\ ]0, +\infty[$ *then the following condition holds:*

$$|W(w, \mathbf{x})| \leq a|w|^2 + b \tag{9}$$

4. $\forall\ \mathbf{x} \in X$, $W(\cdot, \mathbf{x})$ *is continuous.*

5. $W'_{\pm}(w, \mathbf{x}) = \int_Y V'_{\pm}(y, w) d\rho_x(y)$

Proof The first two items are direct consequences of the definition of W and hypothesis 1) and 2) on the loss function V. The fact that W is finite is again a consequence of the hypotheses on V. In fact V is measurable and bounded since

$$|V(y, w)| \leq a|w|^2 + b \quad \forall\ y \in Y,$$

hence we have that V is integrable since $\rho_{\mathbf{x}}(y)$ is finite, direct integration gives the inequality in item three. The continuity of W is straightforward consequence, by remark 4, of covexity and finiteness of $W(\cdot, \mathbf{x})$. Let now consider the fifth item. The existence of the left and right derivatives of $V(y, \cdot)$ and $W(\cdot, \mathbf{x})$ descends from convexity. Furthermore, the stated equality is the consequence of a standard result about the derivative of an integral, under the assumption that in a neighborhood of w the following holds

$$|V'_{\pm}(y, \cdot)| \leq M(y),\ \ M \in L^1(Y, \rho_{\mathbf{x}}). \tag{10}$$

However in our setting this assumption is always fulfilled since the convexity of V implies

$$V'_{-}(y, w - \delta) \leq V'_{\pm}(y, w) \leq V'_{+}(y, w + \delta),$$

and hypothesis 4) on the loss function together with compactness of Y assures that $V'_{\pm}(y, w)$ are bounded functions of y.

■

Using result from Section 3.2 we now evaluate the subgradient $(\partial I_0)(f)$.

Lemma 11 *Consider the functional $I_0 : L^2(X,\nu) \to [0,+\infty[\ \cup\ \{+\infty\}$ defined by*

$$I_0[f] = \int_X W(f(\mathbf{x}),\mathbf{x})d\nu(\mathbf{x}).$$

The following two statements are equivalent.

1. *$f^* \in (\partial I_0)(f)$*
2. *$f^* \in L^2(X,\nu)$ and can be written as*

$$f^*(\mathbf{x}) = \int_Y \alpha(y,\mathbf{x})d\rho_{\mathbf{x}}(y) \tag{11}$$

where

$$\alpha(y,\mathbf{x}) \in (\partial V)(y,f(\mathbf{x})) \quad a.e.[\rho] \tag{12}$$

Proof We first observe that the functional

$$I_0[f] = \int_X W(f(\mathbf{x}),\mathbf{x})d\nu(\mathbf{x})$$

is exactly the Nemetzki functional associated to W. We note that due to inequality (9) above $I_0[f] < +\infty \ \ \forall f \in L^2(X,\nu)$. Then from Proposition 7 we have that I_0 is continuous. Furthermore from Proposition 8

$$f^* \in (\partial I_0)(f) \Leftrightarrow f^*(\mathbf{x}) \in (\partial W)(f(\mathbf{x}),\mathbf{x}), \tag{13}$$

where $f^* \in L^2(X,\nu)$. Now from Section 3.1 we know that

$$(\partial W)(w,\mathbf{x}) = [W'_-(w,\mathbf{x}), W'_+(w,\mathbf{x})]$$

then from (13) it follows that $\exists\ t(\mathbf{x}) \in [0,1]$ s.t.

$$\begin{aligned} f^*(\mathbf{x}) &= t(\mathbf{x})W'_-(f(\mathbf{x}),\mathbf{x}) + (1-t(\mathbf{x}))W'_+(f(\mathbf{x}),\mathbf{x}) = \\ &= \int_Y [t(\mathbf{x})V'_-(y,f(\mathbf{x})) + (1-t(\mathbf{x}))V'_+(y,f(\mathbf{x}))]d\rho_{\mathbf{x}}(y) = \\ &= \int_Y \alpha(\mathbf{x},y)d\rho_{\mathbf{x}}(y), \end{aligned}$$

where

$$\alpha(\mathbf{x},y) = t(\mathbf{x})V'_-(y,f(\mathbf{x})) + (1-t(\mathbf{x}))V'_+(y,f(\mathbf{x})),$$

and

$$\alpha(\mathbf{x},y) \in (\partial V)(y,f(\mathbf{x})).$$

Reversely assume $\alpha(\mathbf{x}, y) \in (\partial V)(y, f(\mathbf{x}))$ $a.e$ $[\rho]$ then

$$f^*(\mathbf{x}) = \int_Y \alpha(\mathbf{x}, y) d\rho_{\mathbf{x}}(y) \ \in \ (\partial W)(f(\mathbf{x}), \mathbf{x}) \ a.e. \ [\nu].$$

The above integral is well defined since inequality (10) ensures that $|\alpha(\mathbf{x}, y)|$ is integrable and hence $f^*(\mathbf{x})$ is finite. Concluding we showed as claimed that given $f \in L^2(X, \nu)$, a function $f^* \in L^2(X, \nu)$ belongs to the subgradient of I_0 at f if and only if

$$f^*(\mathbf{x}) = \int_Y \alpha(\mathbf{x}, y) d\rho_{\mathbf{x}}(y),$$

with

$$\alpha(\mathbf{x}, y) \in (\partial V)(y, f(\mathbf{x})) \ a.e \ [\rho].$$

■

Finally in next theorem we deduce the explicit form of the minimizer of the functional in (7).

Theorem 12 *Given $\lambda > 0$ the problem*

$$\inf_{f \in \mathcal{H}} \{ \int_{X \times Y} V(y, f(\mathbf{x})) d\rho(y, \mathbf{x}) + \lambda \|f\|^2_{\mathcal{H}} \}$$

admits a unique solution $f_\lambda \in \mathcal{H}$ given by

$$f_\lambda = -\frac{1}{2\lambda} \int_{X \times Y} K(\mathbf{s}, \mathbf{x}) \alpha(y, \mathbf{x}) d\rho(y, \mathbf{x})$$

where

$$\alpha(y, \mathbf{x}) \in (\partial V)(f_\lambda(\mathbf{x}), y)$$

where (∂V) is the subgradient of V with respect to its second argument.

Proof Let $\mathcal{J} : \mathcal{H} \to L^2(X, \nu)$ be the canonical inclusion. We note that

$$I[f] = I_0[\mathcal{J}(f)],$$

so that the functional we have to minimize is

$$J[f] = I_0[\mathcal{J}(f)] + \lambda \|f\|^2_{\mathcal{H}}.$$

The above functional is well defined and finite since, as previously noticed, by inequality (9) I_0 is finite on $L^2(X,\nu)$. Now from the linearity of the subgradient (see again Section 3.1) we have that

$$(\partial J)(f) = (\partial I)(f) + 2\lambda f. \tag{14}$$

Moreover from Prop. $III.2.12$ in Ekeland and Turnbull (1983) we know that

$$(\partial I)(f) = \mathcal{J}^*(\partial I_0)(\mathcal{J}(f))$$

where $\mathcal{J}^*$ is the adjoint operator of $\mathcal{J}$. So that to find the subgradient of I we only have to evaluate $\mathcal{J}^*$, that is achieved as follows

$$\begin{aligned} \mathcal{J}^*(f)(\mathbf{s}) &= \langle \mathcal{J}^*(f), K_{\mathbf{s}} \rangle_{\mathcal{H}} = \langle f, \mathcal{J}(K_{\mathbf{s}}) \rangle_2 \\ &= \int_X f(\mathbf{x}) K_{\mathbf{s}}(\mathbf{x}) d\nu(\mathbf{x}) = \int_X K(\mathbf{s},\mathbf{x}) f(\mathbf{x}) d\nu(\mathbf{x}). \end{aligned}$$

Concluding we have that a function f^λ is the unique minimizer of $J[f]$ if and only if

$$0 \in (\partial J)(f^\lambda)$$

(see Section 3.1), that is if and only if

$$f^\lambda = -\frac{1}{2\lambda}\mathcal{J}^*(f^*) \tag{15}$$

with

$$f^* \in (\partial I_0)(\mathcal{J}(f^\lambda))$$

where we used Eq. (14). Finally using Lemma 11 and the explicit form of $\mathcal{J}^*$ we have that the above condition is equivalent to

$$f^\lambda(\mathbf{s}) = -\frac{1}{2\lambda}\int_X K(\mathbf{s},\mathbf{x})\alpha(\mathbf{x},y) d\rho(\mathbf{x},y)$$

where

$$\alpha(\mathbf{x},y) \in (\partial V)(f^\lambda(\mathbf{x}),y) \; a.e. \; [\rho]$$

and

$$\int_Y \alpha(\mathbf{x},y) d\rho_{\mathbf{x}}(y) \in L^2(X,\nu)$$

■

Acknowledgments

L. Rosasco is supported by an INFM fellowship. A. Caponnetto is supported by a PRIN fellowship within the project “Inverse problems in medical imaging”, n. 2002013422. This research has been partially funded by the INFM Project MAIA, the FIRB Project ASTA, and by the EU Project KerMIT.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 686:337–404, 1950.

I. Ekeland and T. Turnbull. *Infinite-dimensional Optimization and Convexity.* Chicago Lectures in Mathematics. The University of Chicago Press, Chicago, 1983.

T. Steinwart. Sparseness of support vector machines. *submitted to IEEE Transactions on Information Theory*, 2002.