

# A NOTE ON THE ROLE OF SQUARED LOSS IN REGRESSION

ANDREA CAPONNETTO

ABSTRACT. In the context of regression, a particularly important role is played by the conditional first moment of the outputs, the so-called regression function. In order to estimate this target function, one usually minimizes a regularized form of the empirical average of a loss function. We show that if no assumption on the underlying probability distribution is available, the squared loss is the only convex loss function whose target is indeed the regression function.

Minimization of penalized loss functionals is a quite common approach in statistical learning. Different loss functions have been proposed in the literature and understanding the pros and cons of different choices is an open issue.

However some choices seem more natural than others. This is the case for example in the regression setting, where one is quite naturally led to search for the first moment of the output conditioned to an input, that is the *regression function*.

It is well known that the regression function attains the minimum of the *expected squared loss*. In fact let

$$(1) \quad V^{sq}(f, y) = (f - y)^2,$$

then assuming that the input-output couples  $(x, y)$  are distributed according to the probability measure  $\rho$ , the expected loss of the estimator  $f(\cdot)$  can be expressed in the form

$$(2) \quad I^{sq}(f) = \int_X \left( \int_{\mathbb{R}} (f(x) - y)^2 d\rho(y|x) \right) d\rho_X(x),$$

where we introduced the marginal and conditional measures,  $\rho_X$  and  $\rho(\cdot|x)$ . Since we are assuming complete knowledge of the probability measure  $\rho$ , regularization is not needed and we can minimize over the space of all the real valued function on the input space. From eq (2) it is clear<sup>1</sup> that this minimization can be accomplished pointwise, in fact the value of the target function  $f^{sq}$  in  $x$  is

$$(3) \quad f^{sq}(x) = \operatorname{argmin}_{w \in \mathbb{R}} \int_{\mathbb{R}} (w - y)^2 d\rho(y|x) \quad \rho_X\text{-a.s. .}$$

Straightforward differentiation with respect to  $w$  of eq (3) gives

$$(4) \quad f^{sq}(x) = \int_{\mathbb{R}} y d\rho(y|x) := f_{\rho}(x) \quad \rho_X\text{-a.s. ,}$$

which shows, as expected, that the regression function is a minimizer of the mean squared loss with no assumption on the probability measure  $\rho$ .

On the other hand, in the learning theory literature it is known that if a boundedness constraint is assumed on the values of the outputs, for example  $\operatorname{supp} \rho_Y \subseteq$

---

*Date:* June 3, 2005.

<sup>1</sup>We will always assume good measurability properties.

$[0, 1]$ ,<sup>2</sup> then the logistic loss shares with the squared loss the above mentioned property. In fact recall the definition of the logistic loss

$$(5) \quad V^{lg}(f, y) = -y \log(f) + (y - 1) \log(1 - f),$$

reasoning as above by differentiation one obtains

$$\begin{aligned} f^{lg}(x) &= \operatorname{argmin}_{w \in \mathbb{R}} \int_{\mathbb{R}} -y \log(w) + (y - 1) \log(1 - w) d\rho(y|x) \\ &= \operatorname{argmin}_{w \in \mathbb{R}} (-f_\rho(x) \log(w) + (f_\rho(x) - 1) \log(1 - w)) = f_\rho(x). \end{aligned}$$

Our purpose is proving that if no restriction is assumed on the probability measure  $\rho$  the squared loss is the only convex loss function (up to constants) such that  $f_\rho$  is always a minimizer of the expected loss.

Let us now pose the problem in a formal way. Consider the loss function  $V(f, y) =: V_y(f)$ ,<sup>3</sup> where  $y$  represents the output and  $f$  the value of the estimator. Assume that  $V_y(\cdot)$  is convex for every  $y$ .

Having in mind the two examples that we have proposed it is clear that in general the property we want to investigate can be analyzed pointwise, that is for fixed  $x$ . Hence we require that for an arbitrary  $x$  and  $\rho(\cdot|x)$ , the following relation must hold

$$\int y d\rho(y|x) \in \operatorname{argmin} \int V_y d\rho(y|x),$$

clearly (recall again the line of reasoning followed for the squared loss case) the condition above is also a sufficient one.

Theorem 1 proves that the previous condition is strong enough to force  $V(f, y)$  being the squared loss. In fact the proof only requires the condition to hold on the class of probability measures  $\sigma := \rho(\cdot|x)$  with support over a couple of real numbers. The Corollary below follows straightforwardly from the previous discussion and Theorem 1.

**Corollary 1.** *Let  $V(f, y) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  be a loss function which is convex in the first argument for every value of the second one. If the expected loss  $I^V(f) := \int V(f(x), y) d\rho(x, y)$  is minimized by the regression function  $f_\rho(x) := \int y d\rho(y|x)$  for every probability measure such that the integrals above are well defined, then  $V(f, y) = A(f - y)^2 + B(y)$  for some constant  $A$  and function  $B(y)$ .*

We now proceed to the proof of Theorem 1.

**Theorem 1.** *Let  $\{V_y\}_{y \in \mathbb{R}}$  a set of finite convex functions on  $\mathbb{R}$ . Assume*

$$(6) \quad \int y d\sigma(y) \in \operatorname{argmin} \int V_y d\sigma(y)$$

for all probability measures  $\sigma$  whose support is a two points set, then  $V_y(f) = A(f - y)^2 + B(y)$  for some constant  $A$  and function  $B$ .

*Proof.* First we prove that  $V_f$  has (two-sided) derivative  $V'_f$  in  $f$ . Consider the probability measure

$$(7) \quad \sigma = (1 - \alpha)\delta_f + \alpha\delta_y.$$

<sup>2</sup>Here  $\rho_Y$  is the marginal probability measure on the output space.

<sup>3</sup>This shorthand notation will simplify expressions in the proof of Theorem 1.

From (6) follows that

$$(8) \quad \int V_y(f + \alpha(y - f))d\sigma(y) \leq \int V_y(f)d\sigma(y).$$

For  $y = f \pm 1$  this inequality can be restated as follows

$$(9) \quad \frac{V_f(\bar{f}) - V_f(f)}{|\bar{f} - f|} \leq V_y(f) - V_y(\bar{f}) + V_f(\bar{f}) - V_f(f),$$

where  $\bar{f} = f \pm \alpha$ . Letting  $\alpha \rightarrow 0^+$  the r.h.s. of the inequality goes to zero due to continuity of convex functions, it follows

$$(10) \quad \partial_- V_f(f) \geq 0 \geq \partial_+ V_f(f)$$

which together with  $\partial_- V_f(f) \leq \partial_+ V_f(f)$  implies  $V'_f(f) = 0$ .

Now define  $D_y$  the subset of  $\mathbb{R}$  where  $V'_y$  exists. It is a known fact ([1] Theorem 25.3) that  $D_y$  contains all but perhaps countably many points (so that in particular  $D_y$  is dense in  $\mathbb{R}$ ), and  $V'_y$  is continuous and non-decreasing relative to  $D_y$ .

Fix an arbitrary sequence of reals  $Y = (y_i)_{i \in \mathbb{N}}$  unbounded both from above and below. The functions  $\{V_y\}_{y \in Y}$  are differentiable on the set  $D_Y := \cap_i D_{y_i}$ , which again contains all but perhaps a countable set of points.

Consider the probability distribution (7) for some  $f \in D_Y$ ,  $y \in Y$  ( $f \neq y$ ) and  $\alpha$  such that  $\bar{f} = f + \alpha(y - f) \in D_Y \cap D_f$ . Since by assumption both  $V'_f(f)$  and  $V'_y(\bar{f})$  exist, condition (6) implies

$$(11) \quad (1 - \alpha)V'_f(\bar{f}) + \alpha V'_y(\bar{f}) = 0,$$

which can be restated as follows

$$(12) \quad (y - f) \frac{V'_f(\bar{f})}{\bar{f} - f} = V'_f(\bar{f}) - V'_y(\bar{f}).$$

Now let  $\alpha$  go to zero, such that  $\bar{f} \rightarrow f$  relative to  $D_Y \cap D_f$ . Due to the known continuity of  $V'_f$  and  $V'_y$  the r.h.s. of the equality tends to  $V'_f(f) - V'_y(f) = -V'_y(f)$ . The l.h.s. must also converge to a limit, so that we obtain

$$(13) \quad (y - f)\phi(f) = -V'_y(f),$$

where  $\phi$  does not depend on  $y$ .

Finally we want to prove that  $\phi(f)$  is constant on  $D_Y$ . Assume  $f_1 < f_2$  in  $D_Y$  s.t.  $\phi(f_1) \neq \phi(f_2)$ . From (13) it descends that for every  $y \in Y$

$$(14) \quad V'_y(f_1) - V'_y(f_2) = y(\phi(f_2) - \phi(f_1)) + C(f_1, f_2).$$

Due to the assumed unboundedness of  $Y$ , we can choose  $y$  large enough in absolute value to make the l.h.s. of the last equality positive. This contradicts the known fact that  $V'_y$  is non-decreasing relative to  $D_y$ .

Since the set  $D_Y$  is dense in  $\mathbb{R}$  and both  $\partial_+ V_y$  and  $\partial_- V_y$  are non-decreasing, the equality

$$(15) \quad 2A(f - y) = V'_y(f)$$

can be extended by continuity to arbitrary  $f$ , and  $V_y$  is differentiable on the whole  $\mathbb{R}$ . Given the level of arbitrariness of  $Y$ , this relation holds for every  $y \in \mathbb{R}$ .

The claim of the theorem follows by integration. □

**Acknowledgments.** We would like to thank T. Poggio and S. Smale for useful discussions and suggestions.

#### REFERENCES

- [1] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

ANDREA CAPONNETTO, CBCL, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, AND D.I.S.I.,  
UNIVERSITÀ DI GENOVA, VIA DODECANESO 35, 16146 GENOVA, ITALY  
*E-mail address:* `caponnet@mit.edu`