Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?

Rémi Gribonval

Should penalized least squares regression be interpreted as Maximum A Posteriori estimation?

Rémi Gribonval, *Senior Member*

Abstract

Penalized least squares regression is often used for signal denoising and inverse problems, and is commonly interpreted in a Bayesian framework as a Maximum A Posteriori (MAP) estimator, the penalty function being the negative logarithm of the prior. For example, the widely used quadratic program (with an ℓ^1 penalty) associated to the LASSO / Basis Pursuit Denoising is very often considered as the MAP under a Laplacian prior. The objective of this paper is to highlight the fact that, while this is *one* possible Bayesian interpretation, there can be other equally acceptable Bayesian interpretations. Therefore, solving a penalized least squares regression problem with penalty $\varphi(x)$ should not necessarily be interpreted as assuming a prior $C \cdot \exp(-\varphi(x))$ and using the MAP estimator. In particular, we show that for *any* prior $p_X(x)$, the conditional mean can be interpreted as a MAP with some prior $C \cdot \exp(-\varphi(x))$. Vice-versa, for *certain* penalties $\varphi(x)$, the solution of the penalized least squares problem is indeed the *conditional mean*, with a certain prior $p_X(x)$. In general we have $p_X(x) \neq C \cdot \exp(-\varphi(x))$.

EDICS*:* SAS-STAT

I. Introduction

Consider the problem of estimating an unknown signal $x \in \mathbb{R}^n$ from a noisy observation $y = x + b$, also known as *denoising*. Given an arbitrary noisy observation y the goal is to estimate the noiseless

Rémi Gribonval is with INRIA, Centre Inria Rennes - Bretagne Atlantique, Campus de Beaulieu, F-35042 Rennes Cedex, Rennes, France. Phone: +33 2 99 84 25 06. Fax: +33 2 99 84 71 71. Email: remi.gribonval@inria.fr.
Rémi Gribonval is a member of the METISS project-team at IRISA, Rennes, France. This work was supported in part by the European Union through the project SMALL (Sparse Models, Algorithms and Learning for Large-Scale data). The project SMALL acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 225913

signal x: in practice, designing a denoising scheme amounts to choosing a function $\psi : \mathbb{R}^n \to \mathbb{R}^n$ which provides estimates of the form $\hat{x} = \psi(y)$. However, unless we specify further what we mean by ”noise” and ”signal”, denoising is a completely ill-posed problem since any pair x, b such that $y = x + b$ can be replaced by $x' = x + z$, $b' = b - z$. Practical denoising schemes hence have to rely on various types of prior information on x and b to design an appropriate denoising function ψ.

A. Bayesian estimation

A standard statistical approach to the denoising problem consists in assuming that x and b are drawn independently at random from known *prior* probability distributions P_X and P_B. Under this *model*, given a cost function $\mathcal{C}(\hat{x}, x)$ that measures the quality of an estimator $\hat{x}$ in comparison to the true quantity to estimate x, the Bayes estimator is defined as an estimator ψ with minimum expected cost:

$$\arg\min_{\psi} \mathbb{E}\left\{\mathcal{C}(\psi(X+B), X)\right\}.$$

For a quadratic cost function $\mathcal{C}(\hat{x}, x) := \|\hat{x} - x\|_2^2$ the Bayes estimator is the conditional mean [5]

$$\psi_\star(y) := \mathbb{E}(X|Y=y). \tag{I.1}$$

Even though this estimator is ”optimal” in the above defined sense, its computation involves a high-dimensional integral and cannot generally be done explicitly. In practice, Monte-Carlo simulations can be used to approximate the integral.

Often more amenable to efficient numerical optimization is the popular Maximum A Posteriori (MAP) criterion, which exploits Bayes rule

$$\begin{aligned} \psi_{MAP}(y) &:= \arg\max_x p(x|y) = \arg\max_x p(y|x)p(x) \\ &= \arg\min_x \left\{-\log p_B(y-x) - \log p_X(x)\right\}. \end{aligned}$$

For white Gaussian noise b, since $p_B(b) \propto \exp(-\|b\|_2^2/2)$, the MAP under the prior p_X can be expressed as

$$\arg\min_x \frac{1}{2}\|y-x\|_2^2 + [-\log p_X(x)]. \tag{I.2}$$

B. Regularization

Optimization problems of the type (I.2) have also been often considered in signal processing without explicit reference to probabilities or priors, under the generic form

$$\arg\min_x \frac{1}{2}\|y-x\|_2^2 + \varphi(x). \tag{I.3}$$

The deterministic objective is to achieve a tradeoff between the data-fidelity term $\|y-x\|_2^2$ and the penalty term $\varphi(x)$, which promotes solutions with certain properties. In particular, when the function φ is non-smooth at the origin, such as $\varphi(x) = |x|^p, 0 < p \leq 1$, the optimum of the criterion (I.3) is known to have few nonzero entries. Regularization with such penalty functions is at the basis of *shrinkage* techniques [3] for signal denoising. More recently, these approaches have become a very popular mean of promoting *sparse* solutions to under-determined or ill-conditioned linear inverse problems $y = \mathbf{A}x + b$, and are now a key tool for compressed sensing [4].

C. Plurality of Bayesian interpretations of regularization

Given the identity of the optimization problems (I.2) and (I.3) when[1] $p_X(x) \propto \exp(-\varphi(x))$, the regularization problem (I.3) is often interpreted as ”solving the MAP under the prior $C \cdot \exp(-\varphi(x))$ (and white Gaussian noise)”. In particular, when $\varphi(x) = \|x\|_1$, a possible interpretation of (I.3) is MAP denoising under a Laplacian prior on x and white Gaussian noise.

The main objective of this paper is to highlight the fact that while the MAP with prior $C \cdot \exp(-\varphi(x))$ is *one* Bayesian interpretation of the estimator (I.3), *there can be other Bayesian interpretations*. We focus on white Gaussian denoising, and we show that for *any* prior $p_X(x)$, the conditional mean can be interpreted as a MAP with some prior $C \cdot \exp(-\varphi(x))$. Vice-versa, for certain functions φ, the estimator (I.3) can equally be interpreted as the *conditional mean*, with a prior $p_X(x)$. In general we do not have $p_X(x) \propto \exp(-\varphi(x))$.

II. Main results

From now on we focus on Gaussian denoising: $B \in \mathbb{R}^n$ is a centered normal Gaussian variable with law $\mathcal{N}(0, \mathbf{I}_n)$ and probability density function (pdf) $p_B(b) \propto \exp(-\|b\|_2^2/2)$. We let $X \in \mathbb{R}^n$ be a random variable independent of B, with law P_X and pdf[2] $p_X(x)$ and $Y = X + B$ be the noisy observation.

In this setting the conditional mean is (see Appendix A)

$$\psi_\star(y) = y + \frac{1}{p_Y(y)} \left[\frac{\partial}{\partial y_i} p_Y(y)\right]_{i=1}^n = y + \nabla \log p_Y(y) \tag{II.1}$$

where $p_Y := p_X \star p_B$ is the pdf[3] of the noisy observation y.

[1]The notation $f(x) \propto g(x)$ means $f(x) = C \cdot g(x)$ for all x, where $C \neq 0$ is some constant independent of x.

[2]For simplicity we consider random variables which admit a pdf.

[3]The pdf p_Y is sometimes refered to as the *evidence* of the observation.

Next we study whether $\psi_\star$ can also be written as the optimum of an optimization problem of the MAP type (I.3), with an appropriate choice of φ. Namely, we investigate when $\psi_\star$ can be identified with the *proximity operator* [2] of a function φ, where we recall the definition

$$\mathrm{prox}_\varphi(y) := \arg\min_{z\in\mathbb{R}^n} \left\{ \frac{1}{2}\|y-z\|_2^2 + \varphi(z) \right\}. \tag{II.2}$$

For smooth φ we have the implicit characterization [2]

$$\mathrm{prox}_\varphi(y) := y - \nabla\varphi[\mathrm{prox}_\varphi(y)], \quad \forall y \in \mathbb{R}^n. \tag{II.3}$$

Comparing with (II.3), we see that if $\psi_\star = \mathrm{prox}_\varphi$ then

$$\nabla\varphi[\psi_\star(y)] = -\nabla \log p_Y(y), \quad \forall y \in \mathbb{R}^n. \tag{II.4}$$

Since $\psi_\star$ is *one-to-one* from $\mathbb{R}^n$ to $\mathrm{Im}\psi_\star$ (see Corollary A.2 in Appendix B), the relation (II.4) characterizes the functions φ such that $\psi_\star = \mathrm{prox}_\varphi$, leading to our theorem.

Theorem II.1. *Consider $Y = X + B$ where $B \sim \mathcal{N}(0, \mathbf{I}_n)$ and $X \sim P_X$ are independent.*

1) *The conditional mean $\psi_\star(\cdot)$ is one-to-one and C^∞ from $\mathbb{R}^n$ onto Im$\psi_\star$. Its reciprocal $\psi_\star^{-1}(\cdot)$: Im$\psi_\star \to \mathbb{R}^n$ is also C^∞.*
2) *We have $\psi_\star$ = prox$_{\varphi_\star}$ where $\varphi_\star : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is defined by:*

$$\begin{aligned} \varphi_\star(x) :=\ & -\tfrac{1}{2}\|\nabla \log p_Y(\psi_\star^{-1}(x))\|_2^2 - \log p_Y[\psi_\star^{-1}(x)], \\ & \textit{for } x \in \textit{Im}\psi_{CM}; \\ \varphi_\star(x) :=\ & +\infty, \\ & \textit{for } x \notin \textit{Im}\psi_\star. \end{aligned} \tag{II.5}$$

3) *If $\tilde\varphi$ satisfies $\psi_\star$ = prox$_{\tilde\varphi}$ then there is a constant $c \in \mathbb{R}$ such that $\tilde\varphi(x) = \varphi_\star(x) + c$ for all $x \in$ Im$\psi_\star$.*
4) *For every $y \in \mathbb{R}^n$, the value $\psi_\star(y)$ = prox$_{\varphi_\star}(y)$ is the* unique local minimum *of the function $\frac{1}{2}\|y-x\|^2 + \varphi_\star(x)$.*

The conditional mean with prior p_X and white Gaussian noise is therefore also the MAP with prior $C \cdot \exp(-\varphi_\star(x))$ and white Gaussian noise.

Remark II.2. Even though the function $x \mapsto \frac{1}{2}\|y-x\|^2 + \varphi_\star(x)$ admits a unique local minimum for any y, the function $\varphi_\star$ defined in (II.5) can be nonconvex, as shown with the following single variable example ($n = 1$). A function $\psi : \mathbb{R} \to \mathbb{R}$ can be written $\psi = \mathrm{prox}_\varphi$ with φ a proper lower semi-continuous function from $\mathbb{R}$ to $\mathbb{R}$ if, and only if, the function ψ is non-expansive and increasing [2]. Here, in the case $n = 1$, $\psi_\star$ is increasing (cf Lemma A.1 in Appendix B), but for certain priors p_X it is expansive (see Remark A.3 in Appendix B): its derivative exceeds one at some point. Since the associated $\varphi_\star$ is C^∞, it is proper and continuous, hence it cannot be convex.

Remark II.3. Caution is in order when interpreting $\psi_\star$ as ”the MAP estimator with prior $\exp(-\varphi_\star(x))$”. This only makes sense if the function $x \mapsto \exp(-\varphi_\star(x))$ is integrable, although in the opposite case some authors refer to the MAP with a ”non-informative prior”.

III. Discussion

For Gaussian priors $X \sim \mathcal{N}(0, \Sigma)$, the conditional mean is the Wiener filter, which is also the MAP and the minimum mean square linear estimator [5], so $\varphi_\star(x) = -\log p_X(x)$.

However, the MAP and the conditional mean are not generically equivalent, so there are choices of p_X (non Gaussian) for which we *do not* have the identity $\varphi_\star(x) = -\log p_X(x)$. Indeed, observe that for any prior $p_X(x)$, the penalty function $\varphi_\star(x)$ defined in Theorem II.1 has the following properties:

- the function $\varphi_\star : \mathrm{Im}\psi_\star \to \mathbb{R}$ is C^∞;
- for any y, the function $x \mapsto \frac{1}{2}\|y - x\|_2^2 + \varphi_\star(x)$ admits a unique local minimum.

Therefore, the identity $\varphi_\star(x) = -\log p_X(x)$ cannot be satisfied if $-\log p_X(x)$ fails to satisfy one of these properties.

For example, generalized Gaussian priors $p_X(x) \propto \exp(-\alpha\|x\|_p^p)$ with $0 < p \leq 1$ are *not smooth* at $x = 0$, hence not in C^∞: for such priors we cannot even have the identity $\varphi_\star(x) = a - b\log p_X(x)$ for any $a, b \in \mathbb{R}$.

One may also wonder whether a reciprocal to Theorem II.1 is possible. Given a penalty function $\varphi(x)$, one can always define $\psi(y) = \mathrm{prox}_\varphi(y)$, and define $q(y) = \psi(y) - y$. However, the main difficulty is to understand when one can write $q(y) = \nabla(p_X \star p_B)(y)$ for some pdf p_X. This is not always possible, for example if $\varphi(x)$ is not sufficiently smooth.

IV. Conclusion and perspectives

We proved that the conditional mean estimator for Gaussian denoising can always be written as a MAP (and that the MAP estimator with certain penalty functions can be interpreted as a conditional mean). These results, in conjunction with Nikolova’s highlighting of model distortions brought by MAP estimation [6], indicate that one should be cautious when interpreting penalized least squares regression scheme in terms of priors:

- *If* the data follows a prior $C \cdot \exp(-\varphi(x))$ and *if* we choose the MAP as a criterion for estimating it, *then* the resulting denoising scheme takes the form of penalized least squares regression with penalty $\varphi(x)$. However, this MAP estimator may have poor denoising performance for this type of data [6].
- In practice, the choice of penalized least squares regression with penalty $\varphi(x)$ is seldomly associated to the *belief* that the data follows the prior $C \cdot \exp(-\varphi(x))$. Instead, it rather stems from the *need* for numerical efficiency and the *empirical observation* that it achieves good denoising performance for the considered class of data.

Given an arbitrary penalty $\varphi(x)$, it remains an open problem to understand for which priors $p_X(x)$ we obtain ”good” denoising performance of penalized least squares regression (for example: performance comparable to the conditional mean).

One can imagine concrete applications of the results presented here for certain priors: in general the conditional mean $\psi_\star(y)$ is *a priori* expressed as an intractable high-dimensional integral; however, if the penalty function $\varphi_\star(x)$ admits a simple expression amenable to efficient numerical optimization (e.g., convex optimization), then the conditional mean can be computed efficiently. Developping such approaches requires a more in depth understanding of the properties of penalty functions $\varphi_\star(x)$ obtained through Theorem II.1. Of particular interest would be the construction of explicit examples where $\varphi_\star(x)$ is ”simple” while p_Y involves an intractable integral.

Another interesting perspective is to obtain alternate statistical interpretations of a larger class of penalized least squares regression estimators (e.g., with non-smooth $\varphi(x)$ such as those leading to sparse estimates). As remarked above, the lack of smoothness makes it impossible to interpret such estimators in terms of a conditional mean, however one may seek interpretations that leave the strict Bayesian framework: for example, one may wish to obtain an interpretation as the optimum of a hybrid Bayesian cost function

$$\min_{\psi} \{\mathbb{E}\mathcal{C}(\psi(X+B),X) + \mathbf{K}(\psi)\}$$

where the term $\mathbf{K}(\cdot)$ forces the function ψ to be in some function class. Eventually, one may also wish to extend theses results to ill-posed linear inverse problems of the type $y = \mathbf{A}x + b$, and to deal with non-Gaussian noise.

V. Acknowledgements

This existence of this paper owes much to several discussions with Mike Davies about the Bayesian ”interpretation” of sparse regularization, as well as intense discussions with Jérôme Idier on the same topic during the second French Spring School of Inverse Problems in Signal Processing, held in the beautiful mediterranean island of Porquerolles in the spring of 2010. The author is very thankful to Mike, Jérôme, and to the organizers of the Spring school for these passionate discussions, and would also like to thank Jean-Christophe Pesquet, who provided his insight on proximity operators, and Patrick Perez, whose comments on a draft version of this paper were precious.

APPENDIX

A. *Proof of the identity* (II.1)

If ψ minimizes the expected square loss, then by the orthogonality relation [5], for any function $\delta : y \mapsto \delta(y)$ we must have $\mathbb{E}\langle \psi(Y) - X, \delta(Y)\rangle = 0$. Hence we obtain the condition

$$\forall \delta, \quad \mathbb{E}\langle \psi(Y) - Y, \delta(Y)\rangle = -\mathbb{E}\langle B, \delta(Y)\rangle.$$

Since $Y = X + B$ has the pdf $p_Y = p_X \star p_B$, we thus require that for any δ

$$\int p_Y(y)\langle \psi(y) - y, \delta(y)\rangle dy = -\mathbb{E}\langle B, \delta(Y)\rangle.$$

Using argument similare to those involved in Stein's risk estimator [7], [1], the right hand side above can be rewritten as follows:

$$\begin{aligned}
-\mathbb{E}\langle B, \delta(X+B)\rangle &= -\mathbb{E}_X \int p_B(b)\langle b, \delta(X+b)\rangle db \\
&= +\mathbb{E}_X \int \langle \nabla p_B(b), \delta(X+b)\rangle db \\
&= \iint p_X(x) \sum_{i=1}^{n} \tfrac{\partial}{\partial b_i} p_B(b) \cdot \delta_i(x+b) dx db \\
&\overset{(a)}{=} \sum_{i=1}^{n} \iint p_X(x) \frac{\partial}{\partial b_i} p_B(y-x) \cdot \delta_i(y) dx dy \\
&= \sum_{i=1}^{n} \int (p_X \star \frac{\partial}{\partial b_i} p_B)(y) \cdot \delta_i(y) dy \\
&= \int \sum_{i=1}^{n} \frac{\partial}{\partial b_i} (p_X \star p_B)(y) \cdot \delta_i(y) dy \\
&= \int \langle \nabla p_Y(y), \delta(y)\rangle dy.
\end{aligned}$$

In (a) we used the change of variable $y = x + b$. We finally obtain the condition: for all y, $p_Y(y)[\psi(y) - y] = \nabla p_Y(y)$. It is easy to check that $p_Y(y)$ cannot vanish, hence this eventually reads $\psi(y) - y = \frac{1}{p_Y(y)} \nabla p_Y(y) = \nabla \log p_Y(y)$.

B. Other technical lemmata

We begin by proving that $\psi_\star$ is always one-to-one.

Lemma A.1. *Denote* $\psi_\star(y) = \left(\psi_\star^i(y)\right)_{i=1}^n$ *where* $\psi_\star^i(\cdot) : \mathbb{R}^n \to \mathbb{R}$ *is scalar valued. The* $n \times n$ *Jacobian matrix* $J[\psi_\star](y) := \left[\frac{\partial}{\partial y_j}\psi_\star^i(y)\right]_{ij}$ *is symmetric positive definite:*

$$\langle v, J[\psi_\star](y) \cdot v\rangle > 0, \quad \forall y \in \mathbb{R}^n, v \neq 0.$$

and satisfies the identity

$$J[\psi_\star](y) = \left[\delta_{ij} + \frac{\partial^2}{\partial y_i \partial y_j} \log p_Y(y)\right]_{ij} = \mathbf{I} + \nabla^2 \log p_Y(y).$$

Proof: For simplicity, we do the proof in the single variable case ($n = 1$). The extension to higher dimension follows the same steps and poses no special difficulty. We indicate the main differences when needed. Since $\psi_\star(y) := y + p'_Y(y)/p_Y(y)$ we have $\psi'(y) = \left(p_Y^2(y) + p''_Y(y)p_Y(y) - [p'_Y(y)]^2\right)/p_Y^2(y)$. Since $n = 1$, what we need to prove is $\psi'(y) > 0$ for all y, or equivalently

$$p_Y^2(y) + p''_Y(y)p_Y(y) - [p'_Y(y)]^2 > 0, \qquad \forall y.$$

Since $p_Y = p_X \star p_B$, $p'_Y = p_X \star p'_B$, $p''_Y = p_X \star p''_B$ and $p_B(b) \propto \cdot e^{-b^2/2}$, we have

$$\begin{aligned} p'_B(b) &\propto e^{-b^2/2} \cdot (-b), \\ p''_B(b) &\propto e^{-b^2/2} \cdot (b^2 - 1) \end{aligned}$$

therefore $p_Y^2(y) + p''_Y(y)p_Y(y) - [p'_Y(y)]^2$ is proportional to

$$\begin{aligned} & \iint p_X(y-b)p_X(y-b') \cdot e^{-(b^2+b'^2)/2} \\ & \cdot \left(1 + \tfrac{b^2-1}{2} + \tfrac{b'^2-1}{2} - bb'\right) db db' \\ = & \iint p_X(y-b)p_X(y-b') \\ & \cdot e^{-(b^2+b'^2)/2} \cdot \tfrac{(b-b')^2}{2} db db' \geq 0 \end{aligned} \tag{A.1}$$

where we used the non-negativity of the integrand[4]. With the change of variable $x = y - b$, $x' = y - b'$, we conclude that $\psi'(y) \geq 0$ with equality only if the function $(x, x') \mapsto p_X(x)p_X(x')$ is identically zero on $\mathbb{R}^2 \backslash \{(x, x), x \in \mathbb{R}\}$. This implies $p_X(x) = 0$ for all x, which is impossible since p_X is a proper pdf. ■

[4]For $n > 1$ the scalar factor $(b - b')^2$ in (A.1) becomes $\langle b - b', v\rangle^2$.

Corollary A.2. *The function $y \mapsto \psi_\star(y)$ is one-to-one from $\mathbb{R}^n$ to Im$\psi_\star$: for any pair $y, y' \in \mathbb{R}^n$, if $\psi_\star(y) = \psi_\star(y')$ then $y = y'$. Moreover, it is C^∞ and its reciprocal is C^∞.*

Proof: We let the reader check that p_Y cannot vanish and is C^∞, hence $\psi_\star$ is C^∞. To prove that $\psi_\star$ is one-to-one, we proceed by contradiction, assuming that $\psi_\star(y) = \psi_\star(y')$ while $y' \neq y$. We define $v := (y'-y)/\|y'-y\|_2$ and the function $f : t \mapsto f(t) := \langle v, \psi_\star(y + tv)\rangle \in \mathbb{R}$. We have $f(0) = f(\|y'-y\|_2)$, and f is smooth, hence its derivative must vanish for some $0 < t < \|y'-y\|_2$. However by Lemma A.1 the derivative is $f'(t) = \langle v, J[\psi_\star](y+tv)\cdot v\rangle > 0$ which yields a contradiction.

■

Remark A.3. The computations done in the proof of Lemma A.1 indicate that for certain choices of the prior p_X we can ensure that $\psi_\star$ is *not* a non-expansive function. We will show it in the single variable case, and similar examples can be built in higher dimensions. By definition, a function $f : \mathbb{R} \to \mathbb{R}$ is non-expansive if $|f(y') - f(y)| \leq |y'-y|$ for all y, y'. If f is differentiable and non-expansive we must have $|f'(y)| \leq 1$ for all y. We prove below that if p_X is symmetric ($\forall x, p_X(-x) = p_X(x)$) and if from some $\varepsilon > 0$ we have $p_X(x) = 0$ for $|x| \leq 1+\varepsilon$, then $\psi'_\star(0) > (1+\varepsilon)^2$.

Proof: It can be checked using the computations done in the proof of Lemma A.1 that

$$\psi'_\star(0) = \frac{\iint p_X(-b)p_X(-b')\cdot e^{-(b^2+b'^2)/2}\cdot\frac{(b-b')^2}{2}dbdb'}{\iint p_X(-b)p_X(-b')\cdot e^{-(b^2+b'^2)/2}dbdb'} > 0.$$

Since p_X is symmetric, easy manipulations show

$$\begin{aligned}\psi'_\star(0) &= \frac{\iint p_X(b)p_X(b')\cdot e^{-(b^2+b'^2)/2}\cdot\frac{b^2+b'^2}{2}dbdb'}{\iint p_X(b)p_X(b')\cdot e^{-(b^2+b'^2)/2}dbdb'}\\ &= \frac{\iint p_X(b)p_X(b')\cdot e^{-(b^2+b'^2)/2}\cdot b^2dbdb'}{\iint p_X(b)p_X(b')\cdot e^{-(b^2+b'^2)/2}dbdb'}\end{aligned}$$

Since $p_X(x) = 0$ for $|x| \leq 1+\varepsilon$ we obtain $\psi'_\star(0) \geq (1+\varepsilon)^2$. ■

C. Proof of Theorem II.1

The fact that $\psi_\star$ is one-to-one and C^∞ with C^∞ reciprocal function was proved in Corollary A.2. We now wish to check that the proximity operator of $\varphi_\star$ defined by (II.5) is indeed $\psi_\star$. The definition of $\varphi_\star(x)$ for $x \notin$ Im$\psi_\star$ ensures that prox$_{\varphi_\star}$ takes its values in Im$\psi_\star$. We let the reader check that a consequence of Lemma A.1 is that the set Im$\psi_\star$ is open. The key point will be to check that there is a *unique* local minimum of $x \mapsto \frac{1}{2}\|y-x\|_2^2 + \varphi_\star(x)$, which is exactly at $\psi_\star(y)$. This will imply in particular that the global minimum $\text{prox}_{\varphi_\star}(y)$ is equal to $\psi_\star(y)$. Denoting x any local mimimum, and u such that $\psi_\star(u) = x$, u must be a local minimum of

$$\begin{aligned} \frac{1}{2}\|y - \psi_\star(u)\|_2^2 + \varphi_\star[\psi_\star(u)] &= \frac{1}{2}\|\psi_\star(u) - y\|_2^2 \\ &\quad -\frac{1}{2}\|\nabla q(u)\|_2^2 - q(u) \end{aligned}$$

(where for the sake of brevity we denoted $q(y) = \nabla \log p_Y(y)$) hence it must satisfy the stationary point equation

$$J[\psi_\star](u) \cdot \big[\psi_\star(u) - y\big] - \nabla^2 q(u) \cdot \nabla q(u) - \nabla q(u) = 0.$$

Using the relation $J[\psi_\star](u) = 1 + \nabla^2 q(u) > 0$ (Lemma A.1) this becomes

$$J[\psi_\star](u) \cdot \big[\psi_\star(u) - y - \nabla q(u)\big] = 0$$

hence $\psi_\star(u) = y + \nabla q(u)$. Since $\psi_\star(u) = u + \nabla q(u)$ we conclude that $u = y$, and therefore $x = \psi_\star(u) = \psi_\star(y)$.

To conclude, assume the function $\tilde{\varphi} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ satisfies $\psi_\star = \text{prox}_{\tilde{\varphi}}$. By (II.4) we must have for all y: $\nabla\tilde{\varphi}[\psi_\star(y)] = -\nabla \log p_Y(y) = \nabla\varphi_\star[\psi_\star(y)]$. In other words, for any $x \in \text{Im}\psi_\star$, $\nabla(\tilde{\varphi} - \varphi_\star)(x) = 0$. Since $\psi_\star$ is a one-to-one mapping of $\mathbb{R}^n$ onto $\text{Im}\psi_\star$, the set $\text{Im}\psi_\star$ is connected hence there must be a constant $C \in \mathbb{R}$ such that for all $x \in \text{Im}\psi_\star$, $\tilde{\varphi}(x) = \varphi_\star(x) + C$.

References

[1] T. Blu and F. Luisier. The SURE-LET approach to image denoising. *IEEE Transactions on Image Processing*, 16(11):2778–2786, 2007.

[2] P. L. Combettes and J.-C. Pesquet. Proximal thresholding algorithm for minimization over orthonormal bases. *SIAM J. Optim.*, 18:1351–1376, 2007.

[3] D. Donoho. Denoising by soft-thresholding. *IEEE Trans. Inform. Theory*, 41(3):613–627, May 1995.

[4] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[5] S. Kay. *Fundamentals of Statistical Signal Processing : Estimation Theory*. Signal Processing. Prentice Hall, 1993.

[6] M. Nikolova. Model distortions in bayesian MAP reconstruction. *Inverse Problems and Imaging*, 1(2):399–422, 2007.

[7] C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.