entropy

MDPI

Article

L1-Minimization Algorithm for Bayesian Online Compressed Sensing

Paulo V. Rossi [1,2,*] and Renato Vicente [1,3]

[1] Latam Experian DataLab , São Paulo-SP 04547-130, Brazil; rvicente@usp.br
[2] Department of General Physics, Institute of Physics, University of São Paulo, São Paulo-SP 05508-090, Brazil
[3] Department of Applied Mathematics, Institute of Mathematics and Statistics, University of São Paulo, São Paulo-SP 05508-090, Brazil
* Correspondence: pvcrossi@if.usp.br; Tel.: +55-11-2608-5104

Received: 3 November 2017; Accepted: 1 December 2017 ; Published: 5 December 2017

Abstract: In this work, we propose a Bayesian online reconstruction algorithm for sparse signals based on Compressed Sensing and inspired by L1-regularization schemes. A previous work has introduced a mean-field approximation for the Bayesian online algorithm and has shown that it is possible to saturate the offline performance in the presence of Gaussian measurement noise when the signal generating distribution is known. Here, we build on these results and show that reconstruction is possible even if prior knowledge about the generation of the signal is limited, by introduction of a Laplace prior and of an extra Kullback–Leibler divergence minimization step for hyper-parameter learning.

Keywords: compressed sensing; L1-minimization; online learning; Bayesian inference

1. Introduction

It has become commonplace to talk about the recent "information explosion" or "data deluge". These expressions refer to a much faster growth in data production compared to all available data storage and to the even more evident divergence between the volume of data produced and the general data processing capacity. This state of things begs for more efficient ways of storing and analyzing data. Natural and man-made signals tend to be compressible, which means their innate redundancies allow them to be expressed as a small number of combinations of its components—they are sparse in some basis. In the last few decades a lot of effort has been directed to the development of compression techniques that allow the sampled data to be rewritten to a reduced number of bits. However, these techniques mean superfluous costs on the gathering and processing of the data. When talking about medical imaging that includes radiation, unnecessary data gathering almost certainly signifies collateral and unwanted health effects in the long run for the patients.

Compressed sensing (CS) is a framework for signal processing that presents itself as a successful resource that has produced many techniques for efficient information extraction [1]. CS utilizes the fact that real-world signals can typically be represented by a small combination of elemental components [2,3] to eliminate the data redundancy directly in its acquisition (in contrast with what occurs in a digital camera, for example, that captures millions of point measurements each time a picture is taken only to discard a large portion of this data immediately after its acquisition through a image compression algorithm). In a standard scenario, CS utilizes the sparsity property to enable the recovery of various signals from much fewer samples of linear measurements than required by the Nyquist–Shannon theorem [4,5]. The Nyquist rate [6,7] is a concept present in virtually all signal acquisition protocols used in consumer electronics and medical imaging devices, among others [2]. Unfortunately, it implies the necessity of a high sampling frequency, which means an especially high demand if we consider the now pervasive high definition registers. Between 2004 and

2006, Candès, Tao and Donoho [4,8,9] showed that sparse signals could be exactly reconstructed with below-the-Nyquist-rate sampling. Moreover (and perhaps what made it so intriguing), the reconstruction could be done in non-adaptive fashion through L1-minimization. These seminal works gave rise to CS [5]. It should be noted that "sampling" here has a twist, though, in CS, a sample of the signal is obtained by calculating its inner product against different test functions. All in all, the popularity of CS in the last few years made it pervasive through various disciplines, from Computer Science to Statistical Physics.

In this work, we consider the scenario of a signal generated by an unknown distribution and its recovery by means of a Bayesian online CS scheme. There is a vast literature on Bayesian schemes for CS (e.g., [10–12]) and their reconstruction have been thoroughly examined, not least by the Statistical Physics community [13–15]. Online reconstruction schemes can be invaluable in situations where real-time inference is needed (e.g., magnetic resonance imaging), when the data set is too large to be entirely stored on a hard disk (e.g., modern simulation science) or in the presence of time-variant signals (e.g., data streaming). Algorithms for Bayesian online learning are well-known in the neural networks community [16–18] and, in many fields of engineering such as optimal control theory, where they are known as Kalman filters [19–21]. In addition, an interesting compromise between online and offline learning is the so-called mini-batch learning [22], which has recently also found its way to the CS scenario [23]. In direct connection with this letter, previous work [24] has already established the possibility of online CS reconstruction, but it suffered from an important limitation, namely the necessity of an exact knowledge about the generating distribution of the signal.

This paper is organized as follows: in Section 2, the basic CS problem is defined, together with two reconstruction methods—the L1-minimization problem called LASSO (Least Absolute Shrinkage and Selection Operator) and the Bayesian scheme. Section 3 introduces the Bayesian online CS framework as presented in [24]. In Section 4, the main contribution of this work, an extension for the online framework in which imperfect knowledge about the signal generating distribution is allowed, is described. Section 5 presents the main results and Section 6 summarizes and discusses the main findings of this work.

2. Problem Setup

Let $\boldsymbol{x}^0 = (x_i^0) \in R^N$ be a real N-dimensional signal where each component is generated by a sparsity-inducing distribution $\phi(x) = (1-\rho)\delta(x) + \rho g(x)$. It is assumed that $0 < \rho < 1$ and that $g(x)$ does not have a finite mass at the origin. Sequential linear projections of the signal with known independent random measurement vectors $\boldsymbol{A}^\mu = (A_i^\mu) \in R^N$, with $A_i^\mu \sim \mathcal{N}(0, N^{-1})$ are obtained at instants $\mu = 1, 2, \ldots$. Let these random projections $u^\mu \equiv \boldsymbol{A}^\mu \cdot \boldsymbol{x}^0$ pass through a (possibly noisy) channel before their result becomes available—this way, the value of each measurement, y^μ, will be distributed according to $P(y^\mu | u^\mu)$ (Figure 1).

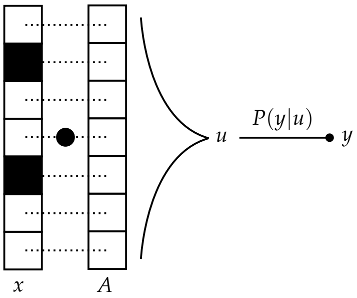

Figure 1. Schematic representation of the measurement process in Compressed Sensing. The sparse signal is projected onto a (known) random measurement vector $\boldsymbol{A}$. The result $u = \boldsymbol{A} \cdot \boldsymbol{x}$ goes through a channel giving rise to the value $y \sim P(y|u)$.

A common goal in Compressed Sensing is to accurately recover the signal $\boldsymbol{x}^0$ based on the knowledge of $D^t = \{(\boldsymbol{A}^1, y^1), (\boldsymbol{A}^2, y^2), \ldots, (\boldsymbol{A}^t, y^t)\}$ and of the functional form of the channel $P(y|\boldsymbol{A} \cdot \boldsymbol{x})$. Since the beginnings of CS, the staple technique in the field [5] has been the LASSO, an optimization problem with an L1-regularization term [25],

$$\hat{\boldsymbol{x}}^t = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda \|\boldsymbol{x}\|_1, \tag{1}$$

where $\boldsymbol{y} = (y^\mu)_{\mu=1}^t$, $\boldsymbol{A}$ is a matrix with all vectors $\boldsymbol{A}^\mu$ as rows and the Lp-norms are defined as $\|\boldsymbol{z}\|_p = (\sum_i |z_i|^p)^{1/p}$. The seminal works [4,8,9] were pivotal in proving that perfect signal reconstruction in absence of noise is possible by means of (1) even in subsampling scenarios (i.e., for $\rho < \alpha < 1$, with $\alpha \equiv t/N$) . The Statistical Physics community is no stranger for these results, with typical behaviour analysis for large systems, where $N \to \infty$ [26,27] being accepted as guarantees for the validity of the method.

More recently, Bayesian techniques for signal reconstruction have also been proposed (e.g., [10,28]). They consist in estimating $\boldsymbol{x}^0$ from the posterior distribution

$$P(\boldsymbol{x}|D^t) = \frac{P(D^t|\boldsymbol{x}) \prod_{i=1}^N \phi(x_i)}{\int d\boldsymbol{x}' \, P(D^t|\boldsymbol{x}') \prod_{i=1}^N \phi(x_i')}. \tag{2}$$

It has been shown [11] that the Bayesian reconstruction typically performs better than the L1-minimization scheme for fewer measurements. In fact, the minimum mean squared error (mmse) estimator $\hat{\boldsymbol{x}}(D^t) = \int d\boldsymbol{x}\, \boldsymbol{x} P(\boldsymbol{x}|D^t)$ is guaranteed [14] to minimize the mean square error mse defined for any possible estimate $\boldsymbol{x}$ as:

$$\mathsf{mse} \equiv N^{-1} \left\langle \|\boldsymbol{x} - \boldsymbol{x}^0\|^2 \right\rangle_{D^t, \boldsymbol{x}^0}. \tag{3}$$

3. Bayesian Online Compressed Sensing

In this work, reconstruction of the signal is done in an online manner. That is, measurements are used one at a time and discarded thereafter (in opposition to the offline/batch scenarios presented above where the whole dataset D^t is available during the entire reconstruction). In addition, we would like to achieve this objective with limited knowledge about the generating distribution $\phi(x)$.

Bayes' Theorem (2) can be easily transformed into an online update recursion [16]. Since all measurements are independent by design, the likelihood can be factored as

$$P(D^t|\boldsymbol{x}) = \prod_{\mu=1}^{t} P(y^\mu|A^\mu \cdot \boldsymbol{x}) = P(y^t|A^t \cdot \boldsymbol{x}) \prod_{\mu=1}^{t-1} P(y^\mu|A^\mu \cdot \boldsymbol{x}). \tag{4}$$

This way, the posterior distribution $P(\boldsymbol{x}|D^t)$ for the signal $\boldsymbol{x}$ after t measurements can be written

$$P(\boldsymbol{x}|D^t) \propto P(y^t|A^t \cdot \boldsymbol{x}) P(\boldsymbol{x}|D^{t-1}). \tag{5}$$

In general, distribution $P(\boldsymbol{x}|D^t)$ can be complicated, so the calculation of $\hat{\boldsymbol{x}}(D^t)$ is potentially difficult. Consider for now the Compressed Sensing scenario where the prior distribution is exactly the same as the known signal generating distribution $\phi(x)$. The $\delta(x)$ factor here adds a singularity to the prior $\phi(x)$, so that the posterior should not approximate a Gaussian distribution even for $t \to \infty$, which was a crucial hypothesis in previous works [16] and would greatly simplify asymptotic calculations. In absence of such a simplification, a previous work [24] introduced the following mean-field approximation:

$$\widetilde{P}(\boldsymbol{x}|D_t) \simeq \prod_{i=1}^{N} \left(\frac{e^{-a_i^t x_i^2/2 + h_i^t x_i} \phi(x_i)}{Z(a_i^t, h_i^t)} \right) \tag{6}$$

for the posterior distribution $P(\boldsymbol{x}|D_t)$ in Label (5) as a device to facilitate marginal computations and (as a consequence) their expected values. In this expression, $\{(a_i^t, h_i^t)\}$ is a set of natural parameters and

$$Z(a_i^t, h_i^t) = \int dx_i e^{-a_i^t x_i^2/2 + h_i^t x_i} \phi(x_i) \tag{7}$$

is the normalization of the marginal distribution $\widetilde{P}_i(x_i|D_t) = \int d\boldsymbol{x}_{\setminus i}\, \widetilde{P}(\boldsymbol{x}|D_t)$, where $d\boldsymbol{x}_{\setminus i}$ denotes integration over all $\{x_j\}_{j \neq i}$. So as not to introduce any biases, we define $h_i^0 = a_i^0 = 0, \forall i$, which results in $\widetilde{P}_i(x_i|D^0 \equiv \varnothing) = \prod_{i=1}^N \phi(x_i)$. The mmse estimate of the signal and its variance can then be readily calculated from the approximate posterior as

$$m_i^t = (\partial/\partial h_i^t) \ln Z(a_i^t, h_i^t), \tag{8}$$

$$v_i^t = (\partial^2/(\partial h_i^t)^2) \ln Z(a_i^t, h_i^t), \tag{9}$$

respectively. Equations (5) and (6) correspond to a sequence of *update* and *projection* steps—the update step adds new information obtained with the most recent measurement t, but transforms the posterior into a possibly intractable distribution; the projection step then returns this distribution to an easier exponential representation with independent components (Figure 2). The full scheme can be summarized as $2N$ update rules for the parameters $\{(a_i^t, h_i^t)\}$ [24]. Defining $\Delta^t \equiv \sum_{i=1}^N A_i^t m_i^{t-1}$ and $\chi^t \equiv \sum_{i=1}^N (A_i^t)^2 v_i^{t-1}$, these rules are

$$a_i^{t+1} = a_i^t - (A_i^{t+1})^2 \frac{\partial^2}{(\partial \Delta^{t+1})^2} \ln \Omega^t(\Delta^{t+1}), \tag{10}$$

$$h_i^{t+1} = h_i^t + A_i^{t+1} \frac{\partial}{\partial \Delta^{t+1}} \ln \Omega^t(\Delta^{t+1}) - m_i^t (A_i^{t+1})^2 \frac{\partial^2}{(\partial \Delta^{t+1})^2} \ln \Omega^t(\Delta^{t+1}), \tag{11}$$

where $\Omega^t(\Delta^{t+1}) = \int \mathcal{D}z P(y^{t+1}|\Delta^{t+1} + \sqrt{\chi^{t+1}}\, z)$ and $\mathcal{D}z = (dz/\sqrt{2\pi}) \exp(-z^2/2)$ is the standard Gaussian measure. It has been shown [24] that this online algorithm has an asymptotic error decay comparable to the offline reconstruction scheme when additive measurement noise is present. In the noiseless scenario, the offline Bayesian algorithm achieves zero error for finite t, while the online version decays exponentially.

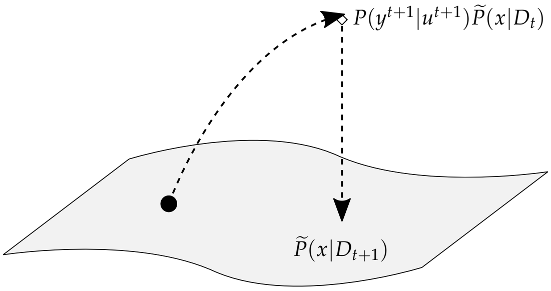
$P(y^{t+1}|u^{t+1})\widetilde{P}(x|D_t)$
$\widetilde{P}(x|D_{t+1})$

Figure 2. Update and projection steps.

4. Mismatched Priors and L1-Minimization Based Reconstruction

A limitation of the framework introduced above is the simple fact that, in a real-world scenario, one does not know the exact parameters of the generating distribution, and possibly not even some of its major characteristics except for the fact that the generated signals are somewhat sparse. Using this algorithm as a canvas, we propose a strategy based on a Bayesian formulation of the L1-regularized problem, relying on the use of the sparsity-inducing Laplace prior

$$\phi_\lambda(x) = (2/\lambda)\exp(-\lambda|x|) \tag{12}$$

and a learning scheme for λ concurrent with the measurements. The use of Laplace-like priors is not new in the literature [29–31], but, to the best of our knowledge, this has not been introduced as part of an online algorithm yet.

We observe that the L1-estimator (1) can also be written in a Bayesian manner as $\hat{\boldsymbol{x}}^t = \int d\boldsymbol{x}\, \boldsymbol{x} \lim_{\beta\to\infty} P_{L1}(\boldsymbol{x}|D^t;\beta,\lambda)$, where

$$P_{L1}(\boldsymbol{x}|D^t;\beta,\lambda) = Z_{\beta,\lambda}^{-1} \exp\left[-\beta\left(\frac{1}{2\sigma_n^2}\sum_{\mu=1}^{t}(y^\mu - \boldsymbol{A}^\mu\cdot\boldsymbol{x})^2 + \lambda\|\boldsymbol{x}\|_1\right)\right], \tag{13}$$

with $Z_{\beta,\lambda}$ a normalization factor dependent on β and λ. This formulation suggests the replacement of the generating distribution $\phi(x)$ in Label (6) for $\phi_\lambda(x)$, effectively altering the projection step of the Bayesian online algorithm to the similar expression

$$\widetilde{P}_\lambda(x|D_t) \simeq \prod_{i=1}^{N}\left(\frac{e^{-a_i^t x_i^2/2 + h_i^t x_i - \lambda|x_i|}}{Z(a_i^t, h_i^t;\lambda)}\right), \tag{14}$$

where $Z(a,h;\lambda) := \int dx\, \exp\left(-ax^2/2 + hx - \lambda|x|\right)$. The first and second moments of this alternate approximate distribution are, respectively,

$$m_i^t = (\partial/\partial h_i^t)\ln Z(a_i^t, h_i^t;\lambda), \tag{15}$$

$$v_i^t = (\partial^2/(\partial h_i^t)^2)\ln Z(a_i^t, h_i^t;\lambda), \tag{16}$$

paralleling Equations (8) and (9). Note that the natural parameters update Equations (10) and (11) remain unchanged, since they are independent of our choice of prior, be it $\phi(x)$, $\phi_\lambda(x)$, or any other.

A common practical issue when making use of the L1-regularization scheme (1) is that the best value of the shrinkage parameter λ is data-dependent and not always clear. Nevertheless, its value is highly influential on the reconstructed model—there is a direct relation between λ and the number of non-zero coefficients in the estimate, which is reflected on the prediction accuracy of the model [32]. Therefore, since it is crucial to estimate accurate values of the shrinkage parameter, usual strategies in offline scenarios include calculation of its best value through cross-validation procedures [25] or straight calculation of an ideal number of non-zero covariates by adding one at a time in a recursive manner [32]. Unfortunately, these strategies are of little use in our present setting, where each measurement has to be used only once in an online manner.

Given any two distributions $q(x)$ and $r(x)$, the Kullback–Leibler divergence [33], $D_{KL}(q||r) = \int dx\, q(x)\ln[q(x)/r(x)]$ is a common measure for their dissimilarity. Also called relative entropy, it has the interesting property that $D_{KL}(q||r) \geq 0$ always and it is zero if and only if $q = r$. Now, whenever distribution $r(x)$ is of the exponential family, i.e., $f(x) \propto \sum_k \xi_k s_k(x)$, minimization of $D_{KL}(q||r)$ with respect to ξ_k is equivalent to matching the moments $\int dx\, q(x)s_k(x)$ and $\int dx\, r(x)s_k(x)$. In particular, for the exact posterior P as defined in the left-hand side of Label (2), minimization of $D_{KL}(P||\widetilde{P}_\lambda)$ with respect to λ after measurement t, i.e., finding λ_t such that

$$\lambda_t = \arg\min_\lambda D_{KL}(P||\widetilde{P}_\lambda) \tag{17}$$

is the same as solving the implicit equation

$$\langle|\boldsymbol{x}|\rangle_P = \langle|\boldsymbol{x}|\rangle_{\widetilde{P}_\lambda} \tag{18}$$

for λ. Here, $\langle |\boldsymbol{x}| \rangle_P$ is the best possible estimate for the true sparsity of the original signal, $Q_0 \equiv \int dx\, |x| \phi(x)$. Our proposal is that λ_t is, at any instant t, an optimal estimate for the shrinkage parameter. Moreover, it is *not static*. With the natural parameters $\{(a_i^t, h_i^t)\}$ evolving, λ_t must change as well to maintain $\langle |\boldsymbol{x}| \rangle_{\widetilde{P}_\lambda} \simeq Q_0$.

Alas, the left-hand side of expression (18) is unknown. To overcome this difficulty, we make the approximation $\langle |\boldsymbol{x}| \rangle_P \simeq |\boldsymbol{m}^t|$. An argument can be made in its defense: as more and more measurements are acquired and the approximate posterior distribution (14) concentrates around the true values, two limits are obtained: (a) $\langle |x| \rangle_{\widetilde{P}_\lambda} \xrightarrow[t\to\infty]{} \langle |x| \rangle_P$ and (b) $\langle |x| \rangle_{\widetilde{P}_\lambda} \xrightarrow[t\to\infty]{} |\langle x \rangle_{\widetilde{P}_\lambda}|$. The first limit has been proved in the large signal size limit $N \to \infty$ in [24] for an exact prior; here, we assume its validity also in the present scenario. As for the second limit, Minkowski inequality guarantees that $\langle |x| \rangle_{\widetilde{P}_\lambda} \leq |\boldsymbol{m}^t|$; with the collapse of the distribution $\widetilde{P}_i(x_i)$ around the true value x_i^0 at $t \to \infty$, no probability mass crosses the $x_i = 0$ axis, so that the equality limit is met. Therefore, these results mean a good estimate for the solution of (18) is given by $\hat{\lambda}_t$ that solves

$$\langle |\boldsymbol{x}| \rangle_{\widetilde{P}_\lambda} \simeq |\boldsymbol{m}^t| \tag{19}$$

for λ, where the right-hand side can be directly calculated from $\widetilde{P}_{\lambda_{t-1}}(\boldsymbol{x}|D^t)$. This implicit expression corresponds to an extra step in the online algorithm for the recovery of the signal, which is shown in full below (Algorithm 1).

Algorithm 1 Online $L1$-based signal recovery for CS.

```
Initialize $P_\lambda(\boldsymbol{x})$                               ▷ $a_i^0, h_i^0 = 0\ (\forall i)$; $\lambda^0 = \lambda$
while $t < t_{\max}$ do
    Obtain new measurement $y^t \sim P(y^t | A^t \cdot x^0)$
    Update $\{(a_i^t, h_i^t)\}$                                  ▷ Equations (10) and (11)
    Find $\hat{\lambda}_t$                                              ▷ Equation (19)
    Update $\lambda_t = (1-\gamma)\lambda_{t-1} + \gamma\hat{\lambda}_t$
    Estimate signal means $m_i^t$ and variances $v_i^t$         ▷ Equations (15) and (16)
end while
return $\boldsymbol{m}^t$
```

5. Results and Discussion

We tested the algorithm on sparse signals generated by two different distributions of the form $\phi(\boldsymbol{x}) = \prod_{i=1}^N \phi(x_i)$, with $\phi(x_i) = (1-\rho)\delta(x_i) + \rho g(x_i)$. The $\delta(x)$ factor induces sparsity by making every component equal to zero with probability $1-\rho$. The non-sparse components could either come from a Gaussian distribution $g(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$ or from a binary distribution $g(x) = (\delta(x+1) + \delta(x-1))/2$. To better evaluate the accuracy losses incurred due to approximation (19), we also tested the algorithm for the ideal situation where Q_0 is known. To guarantee practical usefulness, all results correspond to the standard CS scenario $y^\mu = u^\mu + \xi^\mu$ with Gaussian noise $\xi^\mu \sim \mathcal{N}(0, \sigma_n^2)$, i.e., $P(y|u) = (1/\sqrt{2\pi\sigma_n^2})\exp[-(y-u)^2/2\sigma_n^2]$. Its noiseless counterpart can be obtained by taking the limit $\sigma_n^2 \to 0$, which leads to $P(y|u) = \delta(y-u)$. Figure 3 shows the average mean squared error normalized by $\|\boldsymbol{x}^0\|_2^2$ resulting of simulations of the L1-based online CS algorithm with λ learning. Simulations for finite signal length size N showed decreasing error for increasing N—in order to correct for finite size effects, the curves in Figure 3 were obtained by extrapolating the results for finite $N = 200, 500, 1000$ and 2000 to $N \to \infty$ by means of a quadratic fit. As expected, approximation (19) induces a significant accuracy loss, but, for $\alpha \geq 1$, the reconstruction error decays approximately exponentially, proving the usefulness of the reconstruction scheme.

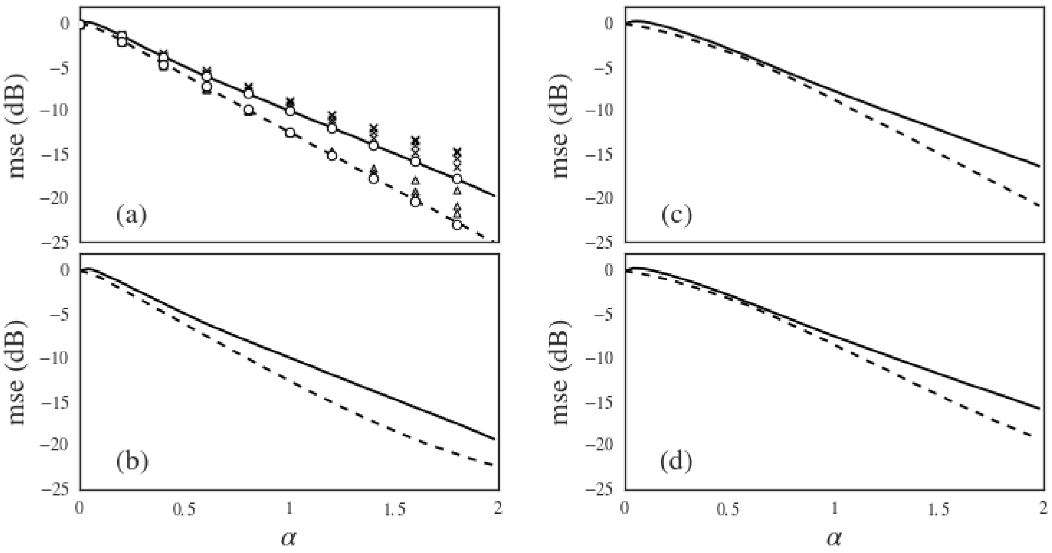

Figure 3. L1-based reconstruction for CS with $\rho = 0.1$ for signals generated by $\phi(\boldsymbol{x}) = \prod_{i=1}^{N}[(1-\rho)\delta(x_i) + \rho g(x_i)]$. (**a**) and (**b**): $g(x) = (1/\sqrt{2\pi})\exp(-x^2/2)$; (**c**) and (**d**): $g(x) = (\delta(x+1) + \delta(x-1))/2$. The top row consists of the noiseless standard CS scenario; the bottom row corresponds to the noisy scenario with $\sigma_n^2 = 10^{-4}$. In all figures, dashed lines were obtained with known Q_0 and full lines with Q_0 estimated through (19). In (**a**), the average error of 100 simulations each of $N = 250, 500, 1000$ and 2000 for chosen values of α are shown as crosses (unknown Q_0) and triangles (known Q_0). The error decreases with increasing N. All lines in all pictures then correspond to the extrapolation of these finite N values to $N \to \infty$ by means of a quadratic fit. All pictures: $\gamma = 10^{-3}$.

In all the examples shown, the solution to Equation (19) was obtained by numerically searching the value $\hat{\lambda}_t$ which best approximates both sides, i.e., by finding $\hat{\lambda}_t = \arg\min_\lambda[\langle|\boldsymbol{x}|\rangle_{\widetilde{P}_\lambda} - |\boldsymbol{m}^t|]^2$. The empirical results prove what could have been a reasonable assumption: Figure 4 shows that in the early stages when only a few measurements have been obtained, even for an *exact* prior, $|\boldsymbol{m}^t|$ is not an excellent approximation of $\langle|\boldsymbol{x}|\rangle_P$. It also shows that the approximation gets progressively better with more measurements. For this reason, we found that the use of a small damping factor γ in the update of λ provides faster and more accurate learning.

Finding an optimal value for the sparsity-inducing parameter λ is not trivial. Ideally, one desires to find λ such that the $\widetilde{P}_\lambda(\boldsymbol{x})$ would represent the actual sparsity of the original signal $\langle|x_i^0|\rangle$. The limit $\lambda \to \infty$ is undesirable, since the soft constraints in $\widetilde{P}_\lambda(\boldsymbol{x})$ due to the finite value of the natural parameters $\{(a_i^t, h_i^t)\}$ would not be enough to prevent the prior distribution to completely dominate the entire approximate posterior, leading all signal estimates to zero. At the same time, consider as a representative example the update Equations (10) and (11) for the noisy standard CS scenario, which can explicitly written as $a_i^{t+1} = a_i^t + (A_i^{t+1})^2/(\sigma_n^2 + \chi^{t+1})$ and $h_i^{t+1} = h_i^t + A_i^{t+1}(y^{t+1} - \Delta^{t+1})/(\sigma_n^2 + \chi^{t+1}) + m_i(A_i^{t+1})^2/(\sigma_n^2 + \chi^{t+1})$. Taking the average of $da_i^\mu = a_i^\mu - a_i^{\mu-1}$ and $dh_i^\mu = h_i^\mu - h_i^{\mu-1}$ over the measurement vectors $\boldsymbol{A}^\mu$ results in

$$\begin{cases} \langle da_i^\mu \rangle_{\boldsymbol{A}^\mu} \simeq N^{-1}/(\sigma_n^2 + \chi^\mu), \\ \langle dh_i^\mu \rangle_{\boldsymbol{A}^\mu} \simeq N^{-1} m_i/(\sigma_n^2 + \chi^\mu), \end{cases} \tag{20}$$

which means that all the natural parameters $a_i^t = \sum_{\mu=1}^{t} da_i^\mu$ typically grow with more and more measurements. Similarly, whenever x_i^0 is different from zero, any parameters $h_i^t = \sum_{\mu=1}^{t} dh_i^\mu$ exhibit a growth in absolute value lead by the m_i factor. For any fixed λ, these considerations mean that the regularizing effect of the prior $\phi_\lambda(x)$ in Label (14) would vanish asymptotically in what would be a typical case of a prior being dominated by the likelihood. It is noteworthy that the introduction of a step for the minimization of the KL-divergence in the Bayesian online algorithm results in estimates $\hat{\lambda}_t$ that typically increase with α as well, in a scale similar to the other natural parameters (Figure 5).

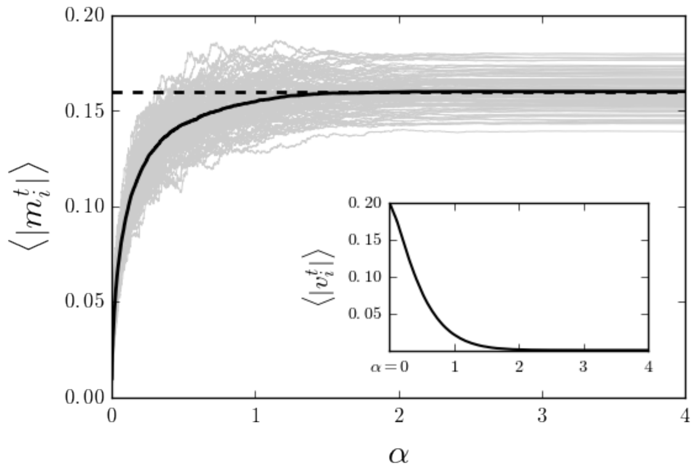

Figure 4. Difference between $\langle |\boldsymbol{x}| \rangle_{\widetilde{P}}$ (from Label (6), where the exact prior is considered) and the true signal sparsity $\langle |x_i^0| \rangle|$. The full line corresponds to the average of 100 simulations of the noisy standard CS scenario with $N = 1000$, $\rho = 0.2$, $\sigma_n^2 = 10^{-4}$ and $g(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$. The dashed line is the true sparsity of the original signal $Q_0 = \int dx^0 \, |x^0| \phi(x^0)$. Inset: Mean variance. Notice that the absolute value approximation gets progressively better with growing α. Indeed, its accuracy matches the diminishing of the posterior distribution's variance.

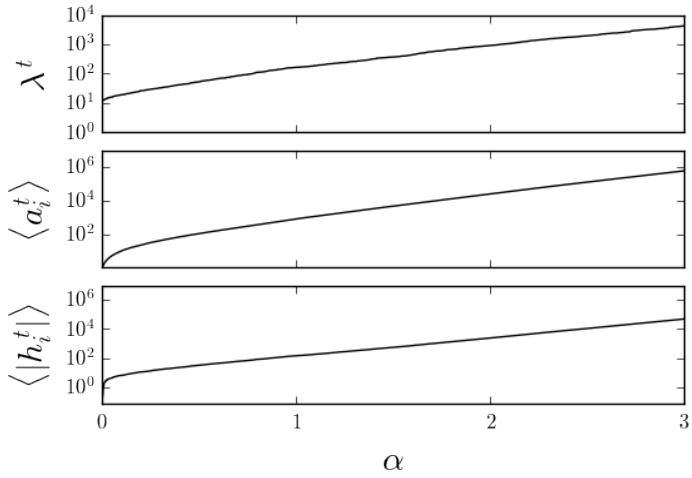

Figure 5. An example of the online CS algorithm where λ has been adjusted after all measurements so that the inferred signal sparsity was equal to the known value Q_0. Note that λ^t, just like the natural parameters $\{(a_i^t, h_i^t)\}$, typically grows exponentially. Here, $N = 1000$, $\rho = 0.1$, $\sigma_n^2 = 0$ and $g(x) = (1/\sqrt{2\pi}) \exp(-x^2/2)$.

6. Conclusions

In this paper, we proposed an online algorithm for Compressed Sensing. Based on an earlier version that expected exact knowledge of the signal generating distribution [24], the present adaptation made use of a double-exponential prior (12) to get rid of this limitation. In order to keep the signal sparsity always in focus during the signal reconstruction, we introduced in this work an extra step where an optimal value for the sparsity-inducing parameter λ can be estimated as a function of $\{(a_i^t, h_i^t)\}$ (i.e., a function of the data D^t). This was achieved through minimization of the KL-divergence between $\widetilde{P}_{\lambda}(x)$ and the (unavailable) full posterior distribution (2), which is approximated by the signal estimate at the previous step.

It is clear that the method is strongly dependent on the validity of the mean-field approximation (6) for the posterior to obtain good reconstruction accuracy. This approximation was introduced in order to simplify the calculations of the moments and to reduce the number of parameters, which should be updated at every step (and, as a consequence, to reduce computational effort and memory requirements). Nevertheless, it is not essential and can be easily replaced by a full Gaussian distribution if needed. In all examples shown here, which made use of signals with fully independent components, the algorithm was shown to give accurate predictions for different signal generating distributions, with and without the presence of additive measurement noise.

Acknowledgments: Paulo V. Rossi would like to thank Yoshiyuki Kabashima for all the discussions that originated this work and Yoshiyuki Kabashima, Andre Manoel and Rafael Calsaverini for all the talks during its production. Renato Vicente would like to thank Julio Stern for many insightful interactions and Paulo V. Rossi and Renato Vicente would like to thank the reviewers for their constructive feedback. This work was partially funded by FAPESP 2014/00792-7 (Paulo V. Rossi). Renato Vicente and Paulo V. Rossi acknowledge the Experian Latam DataLab for supporting this work.

Author Contributions: Paulo V. Rossi and Renato Vicente conceived the algorithm; Paulo V. Rossi performed the simulations; Paulo V. Rossi and Renato Vicente wrote the paper. Both authors have read and approved the fina manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Strohmer, T. Measure What Should be Measured: Progress and Challenges in Compressive Sensing. *IEEE Signal Process. Lett.* **2012**, *19*, 887–893.
2. Candès, E.J.; Wakin, M.B. An Introduction to Compressive Sampling. *IEEE Signal Process. Mag.* **2008**, *25*, 21–30.
3. Holtz, O. Compressive sensing: A paradigm shift in signal processing. *arXiv* **2008**, arXiv:0812.3137.
4. Donoho, D.L. Compressed sensing. *IEEE Trans. Inf. Theory* **2006**, *52*, 1289–1306.
5. Eldar, Y.C.; Kutyniok, G. *Compressed Sensing: Theory and Applications*; Cambridge University Press: Cambridge, UK, 2012.
6. Nyquist, H. Certain topics in telegraph transmission theory. *Trans. AIEE* **1928**, *47*, 617–644.
7. Shannon, C. Communication in the presence of noise. *Proc. Inst. Radio Eng.* **1949**, *37*, 10–21.
8. Candès, E.; Tao, T. Near Optimal Signal Recovery from Random Projections: Universal Encoding Strategies? *IEEE Trans. Inf. Theory* **2006**, *52*, 5406–5425.
9. Candès, E.J.; Romberg, J.; Tao, T. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Trans. Inf. Theory* **2006**, *52*, 489–509.
10. Rangan, S. Generalized approximate message passing for estimation with random linear mixing. In Proceedings of the 2011 IEEE International Symposium on Information Theory Proceedings (ISIT), St. Petersburg, Russia, 31 July–5 August 2011.
11. Krzakala, F.; Mezard, M.; Sausset, F.; Zdeborova, L. Statistical-Physics-Based Reconstruction in Compressed Sensing. *Phys. Rev. X* **2012**, *2*, 021005, doi:10.1103/PhysRevX.2.021005.
12. Tramel, E.W.; Manoel, A.; Caltagirone, F.; Gabrié, M.; Krzakala, F. Inferring sparsity: Compressed sensing using generalized restricted Boltzmann machines. In Proceedings of the 2016 IEEE Information Theory Workshop (ITW), Cambridge, UK, 11–14 September 2016; pp. 265–269.
13. Krzakala, F.; Mezard, M.; Sausset, F.; Sun, Y.; Zdeborova, L. Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices. *J. Stat. Mech. Theory Exp.* **2012**, P08009, doi:10.1088/1742-5468/2012/08/P08009.
14. Xu, Y.; Kabashima, Y.; Zdeborova, L. Bayesian signal reconstruction for 1-bit Compressed Sensing. *J. Stat. Mech. Theory Exp.* **2014**, doi:10.1088/1742-5468/2014/11/P11015.
15. Rangan, S.; Fletcher, A.K.; Goyal, V.K. Asymptotic Analysis of MAP Estimation via the Replica Method and Applications to Compressed Sensing. *IEEE Trans. Inf. Theory* **2012**, *58*, 1902–1923.

16. Opper, M.; Winther, O. Chapter A Bayesian approach to on-line learning. In *On-Line Learning in Neural Networks*; Cambridge University Press: Cambridge, UK, 1998; pp. 363–378.
17. De Oliveira, E.A.; Caticha, N. Inference from aging information. *IEEE Trans. Neural Netw.* **2010**, *21*, 1015–1020.
18. Vicente, R.; Kinouchi, O.; Caticha, N. Statistical mechanics of online learning of drifting concepts: A variational approach. *Mach. Learn.* **1998**, *32*, 179–201.
19. Kalman, R.E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **1960**, *82*, 35–45.
20. Stengel, R.F. *Optimal Control and Estimation*; Courier Corporation: North Chelmsford, MA, USA, 2012.
21. Särkkä, S. *Bayesian Filtering and Smoothing*; Cambridge University Press: Cambridge, UK, 2013; Volume 3.
22. Broderick, T.; Boyd, N.; Wibisono, A.; Wilson, A.C.; Jordan, M.I. Streaming variational bayes. In *Advances in Neural Information Processing Systems*; Curran: Red Hook, NY, USA, 2013; pp. 1727–1735.
23. Manoel, A.; Krzakala, F.; Tramel, E.W.; Zdeborová, L. Streaming Bayesian inference: Theoretical limits and mini-batch approximate message-passing. *arXiv* **2017**, arXiv:1706.00705.
24. Rossi, P.V.; Kabashima, Y.; Inoue, J. Bayesian online compressed sensing. *Phys. Rev. E* **2016**, *94*, 022137, doi:10.1103/PhysRevE.94.022137.
25. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* **1996**, *58*, 267–288.
26. Ganguli, S.; Sompolinsky, H. Statistical Mechanics of Compressed Sensing. *Phys. Rev. Lett.* **2010**, *104*, 188701, doi:10.1103/PhysRevX.2.021005.
27. Kabashima, Y.; Wadayama, T.; Tanaka, T. A typical reconstruction limit for compressed sensing based on Lp-norm minimization. *J. Stat. Mech. Theory Exp.* **2009**, *9*, L09003.
28. Baron, D.; Sarvotham, S.; Baraniuk, R.G. Bayesian Compressive Sensing Via Belief Propagation. *IEEE Trans. Signal Process.* **2010**, *58*, 269–280.
29. Hans, C. Bayesian lasso regression. *Biometrika* **2009**, *96*, 835–845.
30. Park, T.; Casella, G. The bayesian lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686.
31. Figueiredo, M.A. Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2003**, *25*, 1150–1159.
32. Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R. Least angle regression. *Ann. Stat.* **2004**, *32*, 407–499.
33. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.