On the Relationship between Online Optimizers and Recursive Filters

Ömer Deniz Akyıldız
Dept. of Signal Theory and Communications
Universidad Carlos III de Madrid
Leganés, Spain, 28911.
deniz@tsc.uc3m.es

Víctor Elvira
CNRS, CRIStAL
Télécom Lille, Univ. Lille
Villeneuve d'Ascq, France, 59653.
victor.elvira@telecom-lille.fr

Jesus Fernandez-Bes
CIBER, BSICoS Group
I3A, IIS Aragon, University of Zaragoza
Zaragoza, Spain. 50018.
jfbes@unizar.es

Joaquín Miguez
Dept. of Signal Theory and Communications
Universidad Carlos III de Madrid
Leganés, Spain, 28911.
jmiguez@ing.uc3m.es

Abstract

There is a recent interest in the relation between optimization algorithms and the corresponding probabilistic models. Interpreting optimization methods as inference algorithms in probabilistic models provides insight and guides the design of new techniques. This work provides a probabilistic interpretation for incremental proximal methods (IPMs). IPMs are online optimization methods which can be used for minimizing the sum of a large number of component functions. In this paper, we first establish the relationship between the IPM for a linear regression problem and the recursive filtering algorithm for linear-Gaussian models. Interpreting the IPM as a stochastic filter, we discuss the use of a nonlinear recursive filter for optimization as an implementation of an IPM-type algorithm when it is not directly possible to implement the IPM.

1 Introduction

In machine learning and statistics, a common problem of interest is to solve the unconstrained optimization problem

$$\min_{\theta \in \mathbb{R}^d} f(\theta) = \min_{\theta \in \mathbb{R}^d} \sum_{k=1}^{n} f_k(\theta) \tag{1}$$

when n is large. This setting is called large-scale regime since the number of observations is very high to the extent that it rules out the possibility of using classical first or second order optimization algorithms; gradients may be too expensive to evaluate or even to store. Since classical optimization algorithms are not applicable, stochastic optimization algorithms become increasingly popular to tackle the problem given by the Eq. (1). The cornerstone family of algorithms is the well-known stochastic gradient descent algorithms (SGD) which, using a subset of the data, obtain a noisy and unbiased estimate of the true gradient and use it for descent at each step [1]. Arguably, the best-known difficulty is its step-size tuning and there has been an interest in automatic tuning of step sizes, e.g. see [2, 3, 4].

An alternative to SGD methods is called incremental proximal methods (IPMs) [5]. Instead of taking a stochastic gradient step at each iteration, these methods exactly minimize a single component (or

The work was presented at the NIPS workshop on Optimizing the Optimizers. Barcelona, Spain, 2016.

a mini-batch of components) of the cost function by regularizing the solution by the value taken at the previous iteration. However, although straightforward to obtain for the linear case, in general incremental proximal iterations are not easy to obtain for nonlinear regression problems. In that case, every proximal step requires an iterative numerical solution which makes IPMs very unfavorable compared to the SGD. Here, we will consider the simplest instance of IPMs from [5] which will be referred as the IPM henceforth.

In this paper, we provide a probabilistic interpretation of the IPM for large-scale regression problems. The work is similar in spirit to the reinterpretation of optimization algorithms as inference in probabilistic models of [6, 7]. First, we highlight the relationship between the Kalman filter [8] and the IPM and show that these two algorithms essentially result in very similar update rules for the linear case. Then we interpret the extended Kalman filter as a nonlinear version of the IPM.

Related Work. We are not aware of any work in the literature which explicitly highlights the relationship between the IPM and stochastic filters. However the usage of filters for nonlinear optimization problems has been largely addressed in the past. One of the most relevant works in this setting is [9] which provides the case of using extended Kalman filters (EKF) for incremental least-squares problems. In [9], and similar works such as [10], the EKF is viewed as an incremental second-order (Gauss-Newton) method. In [11], the author shows, for a static linear-Gaussian model, that the optimal filter can be seen as a stochastic approximation [1] technique.

There is currently a renascent interest in this connection. In [12], quasi-Newton algorithms are derived as autoregressive filters. Since certain quasi-Newton methods can be seen as the IPM for Hessian matrices at the core, this provides the matrix-variate (more general) version of the interpretation we will provide here for the linear case. In [13], the authors derive the filter for the linear case (they call it probabilistic least-mean squares filter) and provide an efficient approximation of the covariance matrix with a single scalar which then can be seen as the (automatically tuned) step-size of a SGD type method. In [14], the author provides an algorithm called Kalman-based SGD which is equivalent to the algorithm we obtain here in our linear filtering derivation.

While recent work has been mostly focused on the linear case, in this paper, we would like to discuss what nonlinear filtering can provide for solving an *online* optimization problem and vice versa.

2 Incremental Proximal Methods

Incremental proximal methods [5] aim to solve problems of the form (1) by using only a single function at each iteration. In short, the IPM provides solutions to the sequence of problems,

$$\theta_k = \operatorname{prox}_{\lambda, f_k}(\theta_{k-1}) = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} f_k(\theta) + \lambda \|\theta - \theta_{k-1}\|_2^2 \tag{2}$$

where prox denotes the proximal map. Note that, here f_k is actually f_{i_k} where i_k is sampled from the index set $[n] = \{1, \dots, n\}$ uniformly random (as in the case of SGD), but we will slightly abuse the notation and use f_k for simplicity. The IPM generates a sequence $(\theta_k)_{k \geq 0}$ of estimates where each element is a solution of a subproblem. When (2) is analytically solvable, it is argued to be more stable than the SGD and preferable, see [5] for a discussion on convergence.

2.1 IPM for the Linear-Quadratic Cost as a Recursive Filter

Given an output vector $Y \in \mathbb{R}^n$ and inputs $X \in \mathbb{R}^{d \times n}$, the linear regression problem is to fit a vector $\theta \in \mathbb{R}^d$ which satisfies $Y \approx \theta^\top X$. The problem can be formulated as minimizing $f(\theta) = \|Y - \theta^\top X\|_2^2$. Although it can be solved analytically, it is unfeasible to do so if n is large.

Consider $f(\theta) = \sum_{k=1}^n f_k(\theta)$ where $f_k(\theta) = \|y_k - \theta^\top x_k\|_2^2$. For the linear regression case where $f_k(\theta) = \|y_k - \theta^\top x_k\|_2^2$, the incremental proximal iteration will result in the update rule

$$\theta_k = \theta_{k-1} + \frac{x_k (y_k - \theta_{k-1}^\top x_k)}{\lambda + x_k^\top x_k}. \tag{3}$$

The question we want to shed light here is the following: can we obtain (3) as a recursive posterior-mean update in a (Gaussian) probabilistic model? The answer to the question is yes, a similar update rule can be derived using a probabilistic model. Let us consider the following probabilistic model,

$$p(\theta) = \mathcal{N}(\theta; \theta_0, V_0), \qquad p(y_k|\theta) = \mathcal{N}(y_k; \theta^\top x_k, \lambda).$$

where $\mathcal{N}(x; \mu, S)$ denotes the Gaussian distribution defined on x with mean μ and covariance S. Given the data sequence $y_{1:k}$, the posterior distribution $p(\theta|y_{1:k})$ is Gaussian [8]. We denote it as $p(\theta|y_{1:k}) = \mathcal{N}(\theta; \theta_k, V_k)$. The sufficient statistics θ_k and V_k can be computed recursively by,

$$\theta_k = \theta_{k-1} + \frac{V_{k-1}x_k(y_k - \theta_{k-1}^\top x_k)}{\lambda + x_k^\top V_{k-1} x_k}, \tag{4}$$

$$V_k = V_{k-1} - \frac{V_{k-1}x_k x_k^\top V_{k-1}}{\lambda + x_k^\top V_{k-1} x_k}. \tag{5}$$

The relationship between the Eqs. (3) and (4) can be easily seen. At this point, it is also instructive to look at the SGD update for minimizing the linear-quadratic cost which is given by,

$$\theta_k = \theta_{k-1} + \gamma_k x_k(y_k - \theta_{k-1}^\top x_k), \tag{6}$$

where γ_k is the step-size of the algorithm. If we compare (3), (4) and (6), by setting a step size to certain quantities involving x_k and λ, SGD can also be interpreted as a suboptimal filter. An appealing discussion of the connection between stochastic approximation algorithms and optimal filters can be found in [11]. The paper shows (for the linear case) how updates in a form similar to (4) can be seen as valid stochastic approximation methods.

2.2 Extended Recursive Filter as an IPM for Nonlinear Case

Let us consider a nonlinear regression problem where we have $y_k \approx g(x_k, \theta)$ where $g(\cdot, \theta)$ is a nonlinear function of θ. Since x_k's are given (inputs in the machine learning setting), we put $g_k(\theta) := g(x_k, \theta)$ and note that $g_k(\theta) : \mathbb{R}^d \to \mathbb{R}$. The problem of interest is then

$$\min_{\theta \in \mathbb{R}^d} f(\theta) = \min_{\theta \in \mathbb{R}^d} \sum_{k=1}^{n} \|y_k - g_k(\theta)\|_2^2. \tag{7}$$

The incremental proximal iteration for this problem requires to solve

$$\theta_k = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \|y_k - g_k(\theta)\|_2^2 + \lambda \|\theta - \theta_{k-1}\|_2^2$$

at each iteration. In [5], the author assumes this problem is solvable but this is rarely the case. Because of this reason, problems like (7) are usually solved by SGD-type algorithms. To arrive at the extended filtering solution, similarly to the last section, we formulate the probabilistic model,

$$p(\theta) = \mathcal{N}(\theta; \theta_0, V_0), \qquad p(y_k|\theta) = \mathcal{N}(y_k; g_k(\theta), \lambda).$$

Now since the model is nonlinear, the EKF is a natural candidate to use [8]. Let us denote $h_k = \nabla_\theta g_k(\theta_{k-1})$. In this case, extended filtering recursions are given as

$$\theta_k = \theta_{k-1} + \frac{V_{k-1}h_k(y_k - g_k(\theta_{k-1}))}{\lambda + h_k^\top V_{k-1} h_k} \tag{8}$$

and

$$V_k = V_{k-1} - \frac{V_{k-1}h_k h_k^\top V_{k-1}}{\lambda + h_k^\top V_{k-1} h_k}.$$

Note that this is different from a naïve linearization of g_k (i.e. using h_k as the observation model) and then deriving the IPM. In that case, the term $(y_k - g_k(\theta_{k-1}))$ would be replaced by $(y_k - h_k^\top \theta_{k-1})$ which does not result in numerically stable updates.

It is again instructive here to look at the SGD update for nonlinear-quadratic cost functions

$$\theta_k = \theta_{k-1} + \gamma_k h_k(y_k - g_k(\theta_{k-1})),$$

to compare it with (8). From this perspective, SGD can be seen in a similar spirit to extended recursive filters with a hand-tuned covariance.

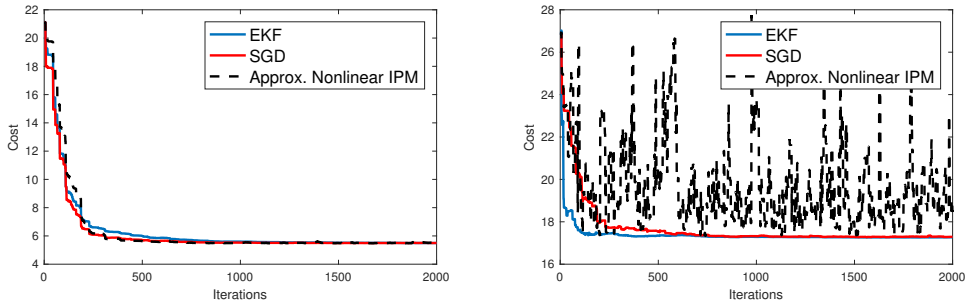

Figure 1: Results on fitting a sigmoid function using EKF, SGD, and approximate nonlinear IPM. On the left, $\lambda_{\text{tr}} = 0.005$ and $\lambda = 1$. On the right, data is much more noisy since $\lambda_{\text{tr}} = 0.05$ and $\lambda = \lambda_{\text{tr}}$. The high noise level, however rightly specified, causes instability for the IPM updates. It is apparent from the experiments that one can safely overestimate the noise level λ and big values of λ is always safer for the IPM. However, the EKF does not suffer from the problem.

3 A Numerical Result and Discussion

As a numerical demonstration, we have compared three algorithms in Fig. 1 on a simple problem of fitting a sigmoid function. The algorithm called approximate nonlinear IPM consists of applying a standard iterative solver for each subiteration since the nonlinear case is not solvable in general. The model used in the experiment is the following model,

$$y_k = g_k(\theta) + \epsilon_k = \frac{1}{1 + \exp(\alpha + \beta^\top x_k)} + \epsilon_k$$

where $\epsilon_k \sim \mathcal{N}(0, \lambda_{\text{tr}})$, and $x_k \in \mathbb{R}^3$ denotes the inputs, and the parameter $\theta = (\alpha, \beta)$ where $\theta \in \mathbb{R}^4$. We set the true value of the parameter λ, denoted with λ_{tr}, to certain values while generating the data and used algorithms with different values of λ (see Fig. 1 for the comments).

We think that the following advantages of a filtering interpretation for online optimization could be beneficial. First, we can estimate λ with standard techniques of maximum likelihood or Bayesian estimation (e.g. by putting an inverse Gamma prior). Second, the EKF would provide error bars for each estimate where neither the IPM nor SGD is able to provide it. The computational burden of updating $d \times d$ matrix V_k at each iteration can be eased using the idea from [13], namely approximating it with a scalar and keep updating only a scalar as a measure of uncertainty.

It should be clear that, in an algorithmic sense, we have not proposed a totally new procedure. What we discussed here is an interpretation of the IPM as a filter, and the EKF as the nonlinear extension of the IPM for general regression problems. Although the use of EKFs for nonlinear regression problems is well-known, the natural choice is often SGD. This is understandable since it is not obvious how the extended filters would behave in a highly nonlinear problem such as training neural networks.

Nevertheless, we think that the filtering interpretation motivated by the IPM can provide guidelines on how to design numerical schemes for more general problems. In machine learning and statistics, problems of the form (1) are popular with nonquadratic and nonlinear cost functions f_k. For many cost functions, a corresponding probabilistic interpretation is possible: for example, the connections between Bregman divergences and exponential families [15] or between Tweedie densities and Beta divergences [16] are well-known. Applying the idea to general cost functions would imply that the solution can be computed recursively with a nonlinear filtering technique. In this direction, there are some initial works in the literature, see e.g. [17], [18]. From the reverse perspective, it would be intriguing to think what other incremental proximal methods proposed in [5], such as incremental subgradient methods, can provide for advanced nonlinear filtering algorithms. Reframing online optimization methods as filters can also be beneficial developing new filtering algorithms and understanding the behaviour of filtering algorithms better.

Acknowledgements. Ö. D. A. and J. M. acknowledge the support of the Office of Naval Research Global (award no. N62909- 15-1-2011) and *Ministerio de Economía y Competitividad* of Spain (project TEC2015-69868-C2-1-R ADVENTURE). J. F. -B.'s work was partially supported by projects TIN2013-41998-R, TEC2014-52289-R, and PRICAM S2013/ICE-2933.

References

[1] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[2] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[3] Tom Schaul, Sixin Zhang, and Yann LeCun. No more pesky learning rates. *ICML (3)*, 28:343–351, 2013.

[4] Maren Mahsereci and Philipp Hennig. Probabilistic line searches for stochastic optimization. In *Advances In Neural Information Processing Systems*, pages 181–189, 2015.

[5] Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical programming*, 129(2):163–195, 2011.

[6] Philipp Hennig, Michael A Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. In *Proc. R. Soc. A*, volume 471, page 20150142. The Royal Society, 2015.

[7] Philipp Hennig. Probabilistic interpretation of linear solvers. *SIAM Journal on Optimization*, 25(1):234–260, 2015.

[8] Simo Särkkä. *Bayesian filtering and smoothing*. Number 3. Cambridge University Press, 2013.

[9] Dimitri P Bertsekas. Incremental least squares methods and the extended kalman filter. *SIAM Journal on Optimization*, 6(3):807–822, 1996.

[10] Bradley M Bell and Frederick W Cathey. The iterated kalman filter update as a gauss-newton method. *IEEE Transactions on Automatic Control*, 38(2):294–297, 1993.

[11] Yu Chi Ho. On the stochastic approximation method and optimal filtering theory. *Journal of Mathematical Analysis and Applications*, 6(1):152 – 154, 1963.

[12] Philipp Hennig and Martin Kiefel. Quasi-newton methods: A new direction. *The Journal of Machine Learning Research*, 14(1):843–865, 2013.

[13] Jesus Fernandez-Bes, Víctor Elvira, and Steven Van Vaerenbergh. A probabilistic least-mean-squares filter. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2199–2203. IEEE, 2015.

[14] Vivak Patel. Kalman-based stochastic gradient method with stop condition and insensitivity to conditioning. *arXiv preprint arXiv:1512.01139*, 2015.

[15] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[16] Y. Kenan Yilmaz and A. Taylan Cemgil. Alpha/beta divergences and tweedie models. *CoRR*, abs/1209.4280, 2012.

[17] Panos Stinis. Stochastic global optimization as a filtering problem. *Journal of Computational Physics*, 231(4):2002–2014, 2012.

[18] Joaquín Míguez, Dan Crisan, and Petar M Djurić. On the convergence of two sequential monte carlo methods for maximum a posteriori sequence estimation and stochastic global optimization. *Statistics and Computing*, 23(1):91–107, 2013.