THE INCREMENTAL PROXIMAL METHOD: A PROBABILISTIC PERSPECTIVE

Ömer Deniz Akyıldız[⋆]*, Víctor Elvira*[†]*, Joaquín Míguez*[⋆]

[⋆]Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés, Spain.
[†]IMT Lille Douai & CRIStAL (UMR CNRS 9189), Villeneuve d'Ascq, France.

ABSTRACT

In this work, we highlight a connection between the incremental proximal method and stochastic filters. We begin by showing that the proximal operators coincide, and hence can be realized with, Bayes updates. We give the explicit form of the updates for the linear regression problem and show that there is a one-to-one correspondence between the proximal operator of the least-squares regression and the Bayes update when the prior and the likelihood are Gaussian. We then carry out this observation to a general sequential setting: We consider the incremental proximal method, which is an algorithm for large-scale optimization, and show that, for a linear-quadratic cost function, it can naturally be realized by the Kalman filter. We then discuss the implications of this idea for nonlinear optimization problems where proximal operators are in general not realizable. In such settings, we argue that the extended Kalman filter can provide a systematic way for the derivation of practical procedures.

***Index Terms*—** Incremental proximal methods, Kalman filtering, stochastic optimization

1. INTRODUCTION

In signal processing and machine learning, it is often of interest to solve unconstrained optimization problems of the form

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

where $f(\theta)$ is the cost function that is built up from the additions of a large number, n, of terms given by the functions f_k. The parameter to be optimized, θ, is a real vector of dimension d.

The setting of Eq. (1) typically arises when one has a large number of independent observations. When n is large, it is not possible to use classical first and second order optimization algorithms, either because gradients are too expensive to compute or Hessian matrices are impossible to store. Stochastic optimization methods have emerged as a powerful solution to this problem. Among many methods, stochastic gradient descent (SGD), proposed in [1], has gained a significant popularity due to its simplicity and superior performance. SGD uses a randomly chosen subset of data to obtain a noisy and unbiased estimate of the true gradient. The well-known difficulty with this procedure is that one has to tune its step-size carefully to prevent divergence and a significant amount of work has been devoted to this topic, e.g. see [2, 3, 4].

A popular alternative to SGD algorithms is the class of incremental proximal methods (IPMs) [5]. These methods utilize *proximal operators* in an online fashion, i.e., they minimize a single or a mini-batch of components of the cost function in Eq. (1) by using a regularizer that depends on the value taken at the previous iteration. Although proximal operators are straightforward to obtain analytically for the linear case, they are not easy to obtain for nonlinear problems. In these cases, every proximal step requires an iterative numerical solver which makes the IPM computationally disadvantegous compared to SGD.

In this paper, we develop a probabilistic interpretation of the IPM for large-scale problems. This approach enables us to obtain purely recursive IPM-type optimizers in the form of approximate filtering algorithms. To attain this goal, we first highlight the relationship between the Kalman filter and the IPM and show that these two algorithms essentially result in very similar update rules for the linear case. Then we discuss how this idea can be extended to nonlinear optimization problems.

Our work is related to an emerging class of algorithms called *probabilistic numerical methods* [6]. These techniques aim at developing probabilistic models of numerical algorithms, which lead to procedures that explicitly tackle the uncertainties inherent to many numerical problems. Also, there is a body of related work on the usage of stochastic filters for nonlinear optimization. In [7] and [8], the extended Kalman filter (EKF) is viewed as an incremental Gauss-Newton method. In [9], the author shows that the optimal filter for a linear-Gaussian model can be seen as a stochastic approximation [1] method. More recently, in [10], quasi-Newton algorithms are derived as autoregressive filters. In [11], the authors derive the filter for linear regression by obtaining a scalar step-size from the posterior covariance. In [12], the author provides a Kalman-based SGD method which is similar to the algorithm in our linear filtering derivation.

Corresponding e-mail: omerdeniz.akyildiz@uc3m.es.

2. PROXIMAL OPERATORS AS BAYES UPDATES

Proximal algorithms have become popular in the signal processing, machine learning, and optimization literature; see [13] for a review from a signal processing perspective and [14] for a thorough review from an optimization perspective. These algorithms utilize *proximal operators* to move towards the minimum of a cost function. A proximal step can be seen as an implicit gradient step [14]. Although they are very general, implementing proximal operators is not straightforward and this is seen as the main limitation of these methods.

Consider the proximal operator of a function f, defined as

$$\text{prox}_{\lambda,f}(\theta_0) = \underset{\theta\in\mathbb{R}^d}{\text{argmin}}\, f(\theta) + \lambda g(\theta,\theta_0),$$

where $g(\theta,\theta_0)$ is a proper distance and $\lambda \in \mathbb{R}_+$ is a regularization parameter. This definition leads to the interpretation of proximal methods as majorization-minimization schemes [14] when θ_0 is replaced by the current estimate θ_{k-1}. Unfortunately proximal operators are not realizable for general f and g. Now, if one considers a probabilistic model[1]

$$p(y|\theta) \propto \exp(-f(\theta)),$$
$$p(\theta|\theta_0,\lambda) \propto \exp(-\lambda g(\theta,\theta_0)),$$

where y is the observation implicit in the cost function f, we can recover the proximal operator as a maximum-a-posteriori (MAP) estimate, i.e.,

$$\text{prox}_{\lambda,f}(\theta_0) = \underset{\theta\in\mathbb{R}^d}{\text{argmax}}\, p(\theta|y,\theta_0,\lambda).$$

Although it seems a straightforward observation, this fact brings important implications. Specifically, it means that the family of Bayesian numerical methods can be used to implement proximal operators. Moreover, instead of aiming at the MAP estimate, one can estimate the posterior pdf $p(\theta|y,\theta_0,\lambda)$, which can be used to quantify the uncertainty of the estimate provided by the optimizer [6], while the proximal operator has no notion of uncertainty over the solution it provides.

Next, we obtain the proximal operator for the linear regression case explicitly. The following derivation relies on the well-known Bayesian interpretation of the ℓ_2 regularizer as a Gaussian prior, see e.g. [15]. Nevertheless, we find it useful to state it in the proximal setting.

Consider the proximal operator for a function $f(\theta) = (y - \mathbf{x}^\top\theta)^2$ with the proximal term $g(\theta,\theta_0) = \|\theta-\theta_0\|^2_{2,V^{-1}}$,

$$\tilde{\theta} = \text{prox}_{\lambda,f}(\theta_0) = \underset{\theta\in\mathbb{R}^d}{\text{argmin}}(y-\mathbf{x}^\top\theta)^2 + \lambda\|\theta-\theta_0\|^2_{2,V^{-1}},$$

where $y \in \mathbb{R}, \mathbf{x} \in \mathbb{R}^d$ and $\theta,\theta_0 \in \mathbb{R}^d$ and $V \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix. Then, $\tilde{\theta}$ is given by,

$$\tilde{\theta} = \theta_0 + \frac{V\mathbf{x}(y-\mathbf{x}^\top\theta_0)}{\lambda+\mathbf{x}^\top V\mathbf{x}}. \tag{2}$$

Next, we consider the probability model

$$p(y|\theta) = \mathcal{N}(y;\mathbf{x}^\top\theta,\lambda), \tag{3}$$
$$p(\theta) = \mathcal{N}(\theta;\theta_0,V), \tag{4}$$

where $\mathcal{N}(\mathbf{z};\mu,S)$ denotes the Gaussian pdf of the vector $\mathbf{z}$ with mean μ and covariance matrix S. Given this model, the posterior pdf can be written as [16]

$$p(\theta|y) = \mathcal{N}(\theta;\tilde{\theta},\tilde{V}), \tag{5}$$

where

$$\tilde{\theta} = \theta_0 + \frac{V\mathbf{x}(y-\mathbf{x}^\top\theta_0)}{\lambda+\mathbf{x}^\top V\mathbf{x}} \tag{6}$$

and

$$\tilde{V} = V - \frac{V\mathbf{x}\mathbf{x}^\top V}{\lambda+\mathbf{x}^\top V\mathbf{x}}. \tag{7}$$

One can see that the mean in Eq. (6) exactly coincides with the update of Eq. (2). As a byproduct of the Bayesian update, we get the covariance matrix in Eq. (7) as a measure of the uncertainty of our estimate.

In the next section, we apply this interpretation to the family of incremental proximal methods in order to get an online and probabilistic optimizer. Perhaps not surprisingly, this algorithm is related to the Kalman filter for the linear case. Then we discuss its extension to nonlinear optimization problems.

3. INCREMENTAL PROXIMAL METHODS

IPMs are a class of algorithms that aim at solving problems of the form of Eq. (1) by using only a single component f_k at each iteration [5]. Given the estimate of the minimum θ_{k-1}, the IPM generates the next estimate by solving the following problem:

$$\theta_k = \text{prox}_{\lambda,f_k}(\theta_{k-1}) = \underset{\theta\in\mathbb{R}^d}{\text{argmin}}\, f_k(\theta) + \lambda\|\theta-\theta_{k-1}\|^2_{2,V^{-1}}, \tag{8}$$

which uses only the most recent "observation". Let us abuse the notation a bit and let f_k stand for f_{i_k} where i_k is sampled uniformly randomly from the index set $[n] = \{1,\dots,n\}$ (the same as for SGD algorithms). The matrix V is usually selected as the identity matrix but other choices, depending on the geometry of the problem, are possible. When the problem in Eq. (8) is solvable, it is argued that it leads to more stable algorithms than the SGD (see [5] for a discussion on convergence).

[1]Throughout the paper, $p(x)$ denotes the probability density function (pdf) of a random variable x. The notation is argument-wise, e.g., $p(y)$ denotes the pdf of another random variable y. The density $p(x|y)$ is the conditional pdf of x given y.

3.1. The IPM as a Kalman filter for linear regression

Given an observation vector $\mathbf{y} \in \mathbb{R}^n$ and a feature matrix $X \in \mathbb{R}^{d\times n}$, the linear regression problem consists in fitting a vector $\theta \in \mathbb{R}^d$ which roughly satisfies $\mathbf{y} \approx X^\top\theta$. Formally, the problem can be framed as the minimization of $f(\theta) = \|\mathbf{y} - X^\top\theta\|_2^2$. When n is large, solving this problem analytically becomes unfeasible. Hence, stochastic optimization methods are often applied. We note that, in this setting, $f(\theta) = \sum_{k=1}^n f_k(\theta)$ and each component f_k can be written as $f_k(\theta) = (y_k - \mathbf{x}_k^\top\theta)^2$ where $y_k \in \mathbb{R}$ is a single observation and $\mathbf{x}_k$ is a column of the feature matrix X. Then, as shown in Section 2, the incremental proximal iteration can be written as

$$\theta_k = \theta_{k-1} + \frac{V\mathbf{x}_k(y_k - \mathbf{x}_k^\top\theta_{k-1})}{\lambda + \mathbf{x}_k^\top V\mathbf{x}_k}. \tag{9}$$

For a proper Bayesian interpretation, we would like to obtain Eq. (9) as a recursive posterior-mean update in a (Gaussian) probabilistic model. In this case, the probabilistic updates yield to a similar, but *different*, algorithm. Notice that, in the probabilistic interpretation of Section 2, V denotes the posterior covariance. However, in Eq. (9), this matrix is kept constant through iterations, as there is no way to update it. In the online optimization literature, some algorithms are proposed to update this matrix, which amounts to updating the proximal term, such as AdaGrad [2]. As we show below, in the probabilistic approach, the matrix V is naturally updated as the posterior covariance matrix.

Let us consider the model

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

Given the data sequence $y_1, \dots, y_k$, the posterior distribution $p(\theta|y_{1:k})$ is Gaussian [16]. We denote it as $p(\theta|y_{1:k}) = \mathcal{N}(\theta; \theta_k, V_k)$. The mean θ_k and covariance V_k can be computed as [17]

$$\theta_k = \theta_{k-1} + \frac{V_{k-1}\mathbf{x}_k(y_k - \mathbf{x}_k^\top\theta_{k-1})}{\lambda + \mathbf{x}_k^\top V_{k-1}\mathbf{x}_k}, \tag{10}$$

$$V_k = V_{k-1} - \frac{V_{k-1}\mathbf{x}_k\mathbf{x}_k^\top V_{k-1}}{\lambda + \mathbf{x}_k^\top V_{k-1}\mathbf{x}_k}. \tag{11}$$

The relationship between the Eqs. (9) and (10) is evident. In the probabilistic approach, we have V_k instead of V, which means that we obtain a way to *update the proximal term* in the Eq. (8) in a principled way and with an intuitive meaning (V_k quantifies the uncertainty in the solution θ_k).

3.2. Extended Kalman filter as an IPM for nonlinear regression

In this section, we consider a nonlinear regression problem. Given observations $\mathbf{y}$, we would like to obtain $y_k \approx g(x_k, \theta)$ where $g(\cdot, \theta)$ is a nonlinear function of θ. Since the x_k's are fixed, we set $g_k(\theta) := g(x_k, \theta)$ for notational conciseness. Note that $g_k : \mathbb{R}^d \to \mathbb{R}$. Then, we would like to solve the problem

$$\min_{\theta\in\mathbb{R}^d} f(\theta) = \min_{\theta\in\mathbb{R}^d} \sum_{k=1}^n (y_k - g_k(\theta))^2. \tag{12}$$

The incremental proximal step for this problem is

$$\theta_k = \operatorname*{argmin}_{\theta\in\mathbb{R}^d} (y_k - g_k(\theta))^2 + \lambda\|\theta - \theta_{k-1}\|_{2,V^{-1}}^2 \tag{13}$$

for each iteration k. Because this proximal step is intractable in general, the typical choice for problems like in Eq. (12) is the SGD. In what follows, we propose the use of EKF recursions as one-step approximations of the realization of the proximal operator.

To this end, let us consider the probabilistic model

$$p(\theta) = \mathcal{N}(\theta; \theta_0, V_0), \qquad p(y_k|\theta) = \mathcal{N}(y_k; g_k(\theta), \lambda). \tag{14}$$

Since the model is nonlinear, using the EKF is a natural way to solve the regression problem. Let us denote $h_k = \nabla_\theta g_k(\theta_{k-1})$. Then, the EKF recursions can be written as,

$$\theta_k = \theta_{k-1} + \frac{V_{k-1}h_k(y_k - g_k(\theta_{k-1}))}{\lambda + h_k^\top V_{k-1}h_k} \tag{15}$$

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 these updates are different from the ones that would be obtained from a naïve linearization of g_k followed by the derivation of the IPM. For that case, we would have the term $(y_k - h_k^\top\theta_{k-1})$, instead of $(y_k - g_k(\theta_{k-1}))$, which does not ensure numerically stable updates.

3.3. Nonstationary optimization

Throughout our discussion, we have kept the prior $p(\theta)$ static, meaning that θ is assumed to be random but not changing with time. While this assumption is convenient when the cost function is not changing, it does not hold in most realistic settings. For this reason, the authors of [3] consider what they call *nonstationary* losses, where the cost function is also changing with time. Tackling such a scenario is trivial from our perspective, as one only needs to modify the algorithm slightly in order to get a dynamic algorithm. In particular, in addition to the update step, one needs to employ a prediction step, according to the assumed dynamics of the parameter. One can model the degree of nonstationarity by modifying the model over θ_k and filtering algorithms extend to such settings very naturally. We leave the detailed investigation of this aspect for future work.

4. NUMERICAL RESULTS

In this section, we investigate two algorithms on a simple problem of fitting a sigmoid function. The first algorithm, which we refer to as *approximate nonlinear IPM*, consists of applying a standard iterative solver for each subiteration since the nonlinear problem of Eq. (13) is not solvable in general. The second algorithm is the EKF, as explained in Section 3.2. The model used in the experiment is of form Eq. (14) with,

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

where $x_k \in \mathbb{R}^{d-1}$ denotes the inputs, and the parameter vector is $\theta = (\alpha, \beta)$ where $\theta \in \mathbb{R}^d$ with $d = 21$. Recall that, by choosing such a model, we aim at solving a problem of form given in Eq. (12). We set the value of the parameter $\lambda = 0.2$ while generating the data and we use the same value in algorithms. Also, the initial value θ_0 of the approximate IPM is set randomly while the proximal matrix is the $d \times d$ identity, denoted $V = I_d$. Similarly, the prior for the EKF is initialized with (θ_0, V_0) where $V_0 = I_d$.

Figure 1(a) shows that the approximate IPM suffers from numerical instability as the parameter estimate θ_k becomes close to the actual minimum. One reason for this instability is that, in the proximal-type algorithms, there is no natural mechanism to reduce the size of the step taken by the algorithm. A natural remedy would be to update the proximal matrix in a way that it dampens the updates as the number of iterations increases, in a similar way to decreasing the step-size of the SGD. The EKF exactly employs this strategy in a natural way. This fact can be seen from Fig. 1(b)-(c) where the diagonal and nondiagonal entries of the covariance matrix V_k are plotted, respectively. It is evident that the entries of this matrix converge to zero, meaning that the update (15) eventually converges to some point in the parameter space.

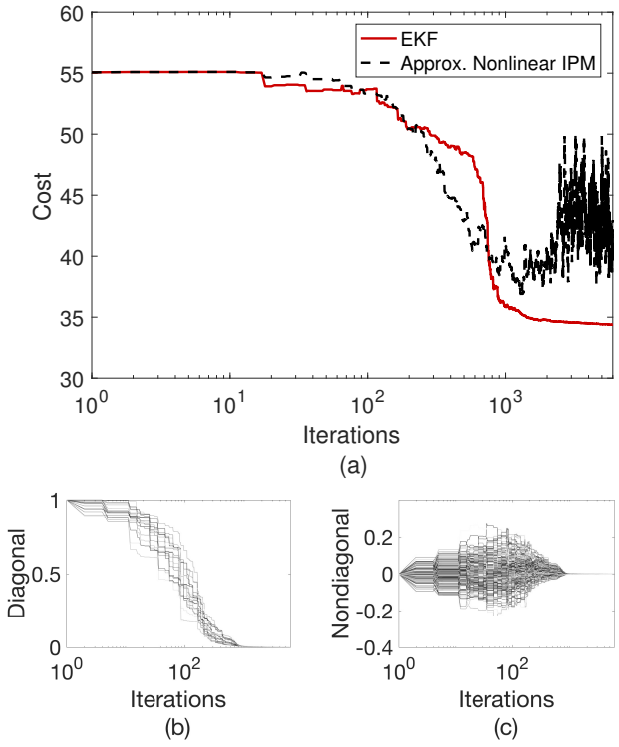

Fig. 1. Results on fitting a sigmoid function using EKF and approximate nonlinear IPM. From (a), it can be seen that the approximate nonlinear IPM proceed towards minimum but suffers from instability while the EKF proceeds in a stable way. From (b)-(c), it can be seen that the entries of the diagonal and nondiagonal parts of the covariance matrix V_k converge to zero which is the reason why the EKF does not suffer from instability.

5. CONCLUSIONS

In this work, we have developed a probabilistic perspective for proximal and incremental proximal methods. We have shown that a probabilistic setting can provide a systematic way to derive algorithms when this is not possible from the classical perspective. In particular, within an online setup, we have argued that the use of filtering algorithms corresponds to employing an IPM-type scheme for optimization. However, filtering algorithms have natural dampening mechanisms for parameter updates, as they refine their uncertainty over the solution iteratively.

This line of work can be pushed forward in a number of different directions. First, different Kalman filters can be used in a similar way to get more advanced optimization schemes for more complicated problems. Among the candidates, the unscented Kalman filter (UKF) and the Ensemble Kalman filter (EnKF) [18] can be useful to tackle high dimensional, possibly time-varying, optimization problems. Second, this approach can be extended beyond quadratic functions by exploiting the relationship between exponential families and Bregman divergences [19]. When the likelihood belongs to the exponential family, the cost function can be expressed in general as a Bregman divergence. In that case, since the Gaussianity assumption is violated, one needs to resort to more complicated numerical algorithms, such as particle filters [20] or other advanced filtering methods.

Acknowledgements

Ö. D. A. and J. M. acknowledge the support of *Ministerio de Economía y Competitividad* of Spain (TEC2015-69868-C2-1-R ADVENTURE), the Office of Naval Research Global (N62909-15-1-2011), and the regional government of Madrid (program CASICAM-CM S2013/ICE-2845). V. E. acknowledges support from the *Agence Nationale de la Recherche* of France under PISCES project (ANR-17-CE40-0031-01).

6. REFERENCES

[1] H. Robbins and S. Monro, "A stochastic approximation method," *Annals of Mathematical Statistics*, vol. 22, pp. 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*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[3] Tom Schaul, Sixin Zhang, and Yann LeCun, "No more pesky learning rates," in *International Conference on Machine Learning*, 2013, pp. 343–351.

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

[5] Dimitri P Bertsekas, "Incremental proximal methods for large scale convex optimization," *Mathematical programming*, vol. 129, no. 2, pp. 163–195, 2011.

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

[7] Dimitri P Bertsekas, "Incremental least squares methods and the Extended Kalman filter," *SIAM Journal on Optimization*, vol. 6, no. 3, pp. 807–822, 1996.

[8] Bradley M Bell and Frederick W Cathey, "The iterated Kalman filter update as a Gauss-Newton method," *IEEE Transactions on Automatic Control*, vol. 38, no. 2, pp. 294–297, 1993.

[9] Yu Chi Ho, "On the stochastic approximation method and optimal filtering theory," *Journal of Mathematical Analysis and Applications*, vol. 6, no. 1, pp. 152 – 154, 1963.

[10] Philipp Hennig and Martin Kiefel, "Quasi-Newton methods: A new direction," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 843–865, 2013.

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

[12] Vivak Patel, "Kalman-based stochastic gradient method with stop condition and insensitivity to conditioning," *SIAM Journal on Optimization*, vol. 26, no. 4, pp. 2620–2648, 2016.

[13] Patrick L Combettes and Jean-Christophe Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*, pp. 185–212. Springer, 2011.

[14] Neal Parikh, Stephen Boyd, et al., "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[15] Rémi Gribonval, "Should penalized least squares regression be interpreted as maximum a posteriori estimation?," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2405–2410, 2011.

[16] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[17] Brian DO Anderson and John B Moore, "Optimal filtering," *Englewood Cliffs, NJ: Pren*, 1979.

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

[19] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, no. Oct, pp. 1705–1749, 2005.

[20] Petar M Djuric, Jayesh H Kotecha, Jianqui Zhang, Yufei Huang, Tadesse Ghirmai, Mónica F Bugallo, and Joaquin Miguez, "Particle filtering," *IEEE signal processing magazine*, vol. 20, no. 5, pp. 19–38, 2003.