On the Relationship between Online Optimizers and Recursive Filters

Ö. Deniz Akyıldız[1], Victor Elvira[2], Jesus Fernandez-Bes[3], Joaquin Miguez[1]

[1]University Carlos III in Madrid, [2]CRIStAL, Telecom-Lille, [3]University of Zaragoza.

Introduction

Interpretation of optimization algorithms as probabilistic inference methods provides insight, paves a way to quantify the uncertainty over the solutions, and can be used to obtain more stable update rules. In this work, we develop a probabilistic insight for a class of online optimization algorithms, called Incremental Proximal Methods (IPMs).

General context. *Online* optimizers randomly sample a data point and update the parameter via some update rule. Stochastic gradient descent (SGD) is the canonical example. This structure is also shared by filtering algorithms although they are derived from an entirely different perspective. We aim to shed light onto this similarity and explore relations.

Notation: $[n] = \{1, \dots, n\}$. At each iteration, we randomly sample $i_k \sim [n]$ and choose f_{i_k} to optimize. We abuse the notation and denote these functions with $f_k := f_{i_k}$.

Incremental Proximal Methods

In machine learning, it is of crucial interest to solve unconstrained optimization problems of the following form,

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{N} f_i(\theta) \tag{1}$$

SGD chooses f_k randomly and takes a (noisy) gradient step. Another class of algorithms which can be used here is called Incremental Proximal Methods (IPMs).

The IPM [1] minimises the cost (1) in the following way,

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

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 solving,

$$\min_{\theta \in \mathbb{R}^d} f(\theta) = \min_{\theta \in \mathbb{R}^d} \|Y - \theta^\top X\|_2^2$$

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

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

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. We formulate,

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

Given the data sequence $y_{1:k}$, the posterior distribution $p(\theta|y_{1:k})$ is Gaussian. 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,

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

and

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

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,

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

Extended Recursive Filter as an IPM for Nonlinear Case

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.

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. We denote $h_k = \nabla_\theta g_k(\theta_{k-1})$ and the EKF recursions are given by,

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

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 naive 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 $(y_k - h_k^\top \theta_{k-1})$.

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.

A Numerical Result and Discussion

We compare the *approximate* IPM, EKF, and SGD on a simple problem of fitting a sigmoid function. The model used in the experiment is the following model,

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

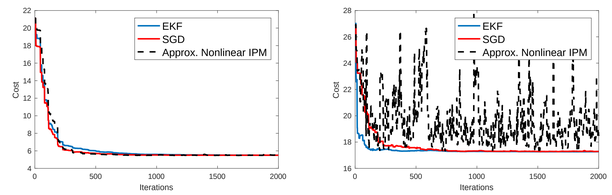

Figure 1: Results on fitting a sigmoid function using EKF, SGD, and approximate nonlinear IPM. On the left, $\lambda_{\text{tr}} = 0.005$ and $\lambda = 1$. On the right, data is much more noisy since $\lambda_{\text{tr}} = 0.05$ and $\lambda = \lambda_{\text{tr}}$. The high noise level, however rightly 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.

References

1. Bertsekas, Dimitri P. "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey." *Optimization for Machine Learning* 2010 (2011): 1-38.

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.