Statistical Filtering for Optimization

Direct, Stochastic Analogues to Deterministic Optimization Methods

Vivak Patel

University of Chicago
OMS 2017 Havana, Cuba

December 16, 2017

Overview

Preliminary presentation of a novel framework for directly applying deterministic solvers for stochastic optimization problems.

These solvers outperform state-of-the-art methods in Machine Learning, Statistics and Applied Mathematics.

Outline

Motivation. Three Examples.

Current Methods. Stochastic and Bayesian Optimization.

Our Insight. Statistical Filtering over Function Spaces.

Experiments. GLMM Example.

Motivation: Empirical Risk Minimization

We have a collection of outcomes and features $\{(y_i, X_i) : i = 1, \ldots, N\}$, and we want to identify $f(X, \beta)$ that returns y.

To identify β, we minimize the empirical risk

$$\min_{\beta} \sum_{i=1}^{N} l(Y_i, f(X_i, \beta)), \tag{1}$$

where $l \geq 0$ is called a loss function.

Motivation: Empirical Risk Minimization

When N is large, the empirical risk minimization problem is solved using stochastic methods such as Stochastic Gradient Descent, Stochastic Quasi Newton and Natural Gradient Descent (i.e., the Kalman Filter) (Bottou et al., 2016).

Motivation: Optimal Inventory Control

Consider the problem of planning production (P_t) and inventory (I_t) for a demand (S_t) over time $t = 0, 1 \ldots, T$ related by $I_{t+1} = I_t + P_t - S_t$.

Suppose S_t are i.i.d. Poisson Random Variables with mean λ.

Suppose a shortage or surplus in inventory is penalized by a cost $c_1 I_t^2$ and production is penalized by $c_2 P_t$.

Motivation: Optimal Inventory Control

The production levels and initial inventory that minimize the cost are given by

$$\begin{aligned} \min_{P_0,\ldots,P_{T-1},I_0} \quad & \mathbb{E}\left[\sum_{t=0}^{T} c_1 I_t^2 + \sum_{t=0}^{T-1} c_2 P_t\right] \\ \text{s.t.} \quad & I_{t+1} = I_t + P_t - S_t,\ t = 0,\ldots,T-1 \\ & P_t \geq 0,\ t = 0,\ldots,T-1. \end{aligned} \tag{2}$$

Complex variants of this problem are solved using Sample Average Approximations (Shapiro et al., 2009).

Motivation: Estimating GLMMs

Consider 500 patients treated by 50 doctors for the same disease. Given patient demographic information, we want to know the probability of being cured.

Suppose we do not care about our set of 50 physicians, but we want to acknowledge the impact of a random physician in our model.

Motivation: Estimating GLMMs

The probability of being cured is then

$$p = \frac{1}{1 + \exp(-\beta' X + \sigma Z)}. \tag{3}$$

The likelihood is

$$\mathcal{L}(\beta, \sigma) = \mathbb{E}\left[\prod_{i=1}^{500} p_i^{y_i}(1 - p_i)^{1-y_i}\right], \tag{4}$$

β and σ are estimated either by approximating the integral or by approximating the integrand (Tuerlinckx et al., 2006).

Motivation: Abstraction

Problem Abstraction:

$$\min_{x} \mathbb{E}\left[f(x, W)\right], \tag{5}$$

where W is a random variable that can be sampled; and x may be subject to some convex constraints or equality constraints.

We also require $\nabla\mathbb{E}[f(x, W)] = \mathbb{E}[\nabla f(x, W)]$.

Methods: Overview

Incremental Estimators. Stochastic Gradient Descent, Kalman-based SGD, Stochastic QN, AdaGrad.

Stochastic Optimization. Stochastic Average Gradient, Stochastic Variance Reduced Gradient, Stochastic Dual Coordinate Descent.

Approximations. Finite Sample Surrogate, Laplace's Approximation.

Bayesian Optimization. Gaussian Process Surrogate for Objective Function.

Methods: Goals

Our per-Iterate Sample Size Cannot Grow. Commonly believed to be prevent higher-order convergence rates for continuous random variable.

We want to solve the exact problem. Our surrogate must converge to the true function.

Mimic Deterministic Methods. We want objective function values, gradient norm values, line-search/trust-region, & stopping criteria.

Insight: Strategy

Build Surrogates. For both the objective, the gradient and (possibly) the Hessian.

Converge Along Iterates. Like the Dennis-Moré condition, we need our surrogate to converge along the iterates.

Compactly Propagate Uncertainty. Avoids increasing samples or pre-scheduling step sizes for convergence.

All three of these goals can be accomplished by statistical filters. In particular, we must shift from filtering over parameters ($\mathbb{R}^d$) to statistical filtering over function spaces.

Insight: Statistical Filter

A SF estimates a Hidden Markov Model (Simon, 2006).

A HMM requires defining ``hidden states'' and relationships between hidden states.

A HMM requires defining ``observed states'' and their relationships to hidden states.

Insight: HMM for Functions

Let $\{x_0, x_1, \ldots\}$ be a sequence.

Hidden States. $\mathbb{E}[f(x_i, W)]$ and $\nabla\mathbb{E}[f(x_i, W)]$. Possibly include the Hessian.

Observed States. Let $\{W, W_0, W_1, \ldots\}$ be i.i.d. Then, $f(x_i, W_i)$ and $\nabla f(x_i, W_i)$.

Relationships?

We use Taylor's theorem and Lipschitz Continuity of the gradient (or Hessian) to relate the hidden states, and correct for the systematic error by inflating the propagated covariance.

Insight: SF Along Iterate Path

Under the appropriate setting, using the Kalman Filter a filter over the objective and gradient with a sequence $\{x_0, x_1, \ldots\}$ converging to x^*, the estimates of the objective and gradient, $\widehat{F}_i(x_i)$ and $\widehat{G}_i(x_i)$, converge to $\mathbb{E}[f(x^*, W)]$ and $\nabla\mathbb{E}[f(x^*, W)]$ in L^2 and in probability.

Insight: Function Estimation & Optimization

Given $\widehat{F}_{i-1}(x_{i-1})$ and $\widehat{G}_{i-1}(x_{i-1})$, use the statistical filter to define $\widehat{F}_i(x)$.

Generate x_i by solving the surrogate (up to one iteration)

$$\min_x \widehat{F}_i(x), \tag{6}$$

using a deterministic method.

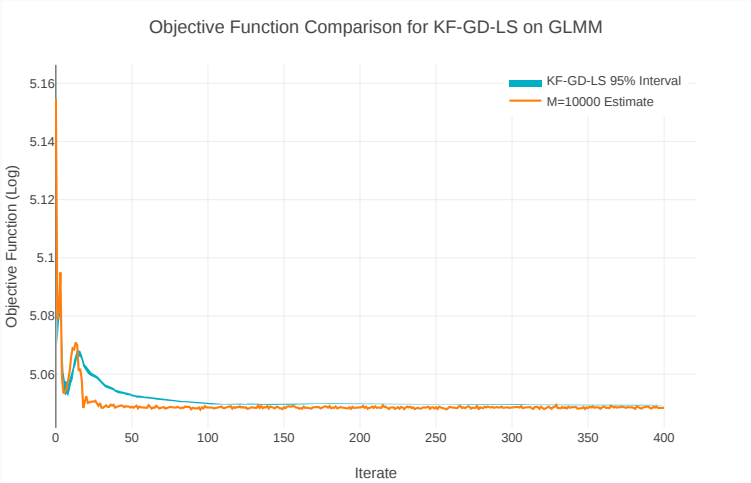
Experiment: Setup

Consider the GLMM problem with 250 patients with a disease treated by 90 physicians. There are 5 fixed effects and 1 random effect that must be estimated.

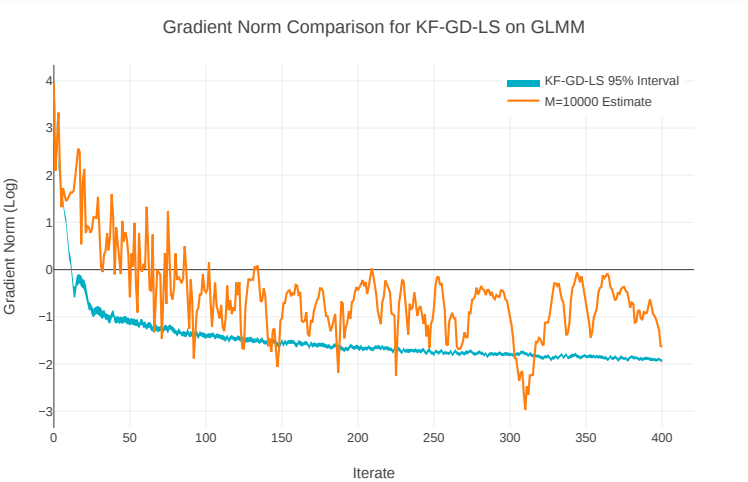
Because there are such few patients per physician, inference techniques described by Tuerlinckx et al. (2006) are known to fail.

We apply a Kalman Filter to generate the surrogates, and solve the surrogate optimization problem with Gradient Descent and Line Search.

Experiment: Objective Function



Experiment: Gradient Norm



Summary

We presented a framework that builds compactly representable surrogate functions of $\mathbb{E}[f(x, W)]$ that can be minimized using deterministic solvers.

The resulting stochastic optimization solver inherits the nice features of deterministic solvers.

The resulting solvers outperform current state-of-the-art methods on the experiments that we have run.

References

Bottou, L., F. E. Curtis, and J. Nocedal
2016. Optimization methods for large-scale machine learning. arXiv preprint arXiv:1606.04838.

Shapiro, A., D. Dentcheva, and A. Ruszczyński
2009. Lectures on stochastic programming: modeling and theory. SIAM.

Simon, D.
2006. Optimal state estimation: Kalman, H infinity, and nonlinear approaches. John Wiley & Sons.

Tuerlinckx, F., F. Rijmen, G. Verbeke, and P. Boeck
2006. Statistical inference in generalized linear mixed models: A review. British Journal of Mathematical and Statistical Psychology, 59(2):225--255.

Acknowledgements

Mihai Anitescu for his general guidance.

NSF RTG 1547396 for its financial support.

Thank You

www.vivakpatel.org