DIRECT, STOCHASTIC ANALOGUES TO DETERMINISTIC OPTIMIZATION METHODS USING STATISTICAL FILTERS *

VIVAK PATEL †

Abstract. Stochastic optimization—those problems that involve random variables—is a fundamental challenge in many disciplines. Unfortunately, current solvers for stochastic optimization restrictively require finiteness by either replacing the original problem with a sample average surrogate, or by having complete knowledge of a finite population. To help alleviate this restriction, we state a general, novel framework that generates practical, robust numerical methods to solve the actual stochastic optimization problem iteratively. Our key insight is to treat the objective and its gradient as a sequential estimation problem that is solved by integrating statistical filters and deterministic optimizers. We demonstrate the framework by generating a Kalman Filtering-based gradient descent method with line search or trust region to solve a challenging stochastic optimization problem in statistics.

Key words. Stochastic Optimization, Statistical Filtering, Direct Analogues, Convergent Surrogate Models

1. Introduction. From statistics to control, a ubiquitous optimization problem is

$$\min_{x \in \mathbb{R}^d} \mathbb{E}\left[f(x, W)\right] \tag{1.1}$$

where W is an $\mathbb{R}^p$-valued random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$; $\mathbb{E}\left[\cdot\right]$ denotes the expectation operator with respect to the probability space; and $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ is a continuously differentiable functional and is sufficiently regular such that $\nabla \mathbb{E}\left[f(x, W)\right] = \mathbb{E}\left[\nabla f(x, W)\right]$. Currently, this optimization problem is addressed using three broad approaches with their unique advantages and challenges.

1. In the sample average approximation (SAA) approach, the objective function is replaced with an empirical mean generated from an independent sample of the random variable [8]. The resulting objective function is deterministic and can be solved with mature deterministic solvers. However, for more complex variants of (1.1), the SAA objective function with only, say, a thousand samples becomes prohibitively expensive to manipulate per iteration [4].

2. In the Bayesian Optimization approach, the objective function is nonparametrically estimated using a Gaussian process prior that is updated at each iteration using an evaluation of the function f at some point in $\mathbb{R}^d$. [3]. Under mild conditions, the estimated objective function converges to the actual objective function in the limit. However, just for modeling the objective function, the estimated objective function update requires inverting a dense matrix whose dimension is equal to the iteration number, which becomes prohibitively expensive as the number of iterations grows and especially if the gradient is modeled using the same procedure.

3. In the Incremental Estimation approach (e.g. Stochastic Gradient Descent), the objective function is ignored and sampled gradient information is used exclusively to perform the optimization [2]. The incremental estimation approach is distinguished by its inexpensive per iteration costs and simplicity of use. However, while the iterates can be shown to converge to a stationary point, such methods are extremely difficult to tune and they do not provide stopping criteria [5, 7] (see [6] for an exception).

*The author is supported by NSF RTG Award 1547396.

†University of Chicago, (vp314@uchicago.edu, www.vivakpatel.org).

Despite the challenges of these three approaches, each has highly desirable advantages such as the ability to use deterministic solvers, the convergence of surrogates to the original problem, and inexpensive per iteration costs. In an ideal situation, there would be an optimization methodology that retains all of these benefits. In this work, we introduce such a framework that generates iterative solvers for (1.1). Our key insight is to treat the objective function and the gradient function as a Hidden Markov Model *over function spaces*, and to integrate statistical filters and deterministic iterative solvers to practically estimate and minimize (1.1). The remainder of this work describes this methodology.

2. Principles of Statistical Filtering. A statistical filter is an estimation methodology for a Hidden Markov Model (HMM) [10, 9]. A basic HMM is defined by three components: an initial distribution, μ, a Markov transition kernel, p, and an observation distribution, q. The first two components define a Markov Chain, $\{X_0, X_1, X_2, \ldots\} \subset \mathbb{R}^h$, where $X_0 \sim \mu$ and $\mathbb{P}\left[X_{i+1} \in A \middle| X_i\right] = p(X_i, A)$ for a measurable set A. The third component defines a sequence of observations $\{Y_1, Y_2, \ldots\} \subset \mathbb{R}^b$ such that $\mathbb{P}\left[Y_i \in B \middle| X_i\right] = q(B|X_i)$ for a measurable set B.

The sequential estimation of a HMM follows from Bayes' Rule. In particular, given a state X_i and an observation Y_{i+1}, the distribution of X_{i+1} is

$$\mathbb{P}\left[X_{i+1} \middle| X_i, Y_{i+1}\right] \propto \mathbb{P}\left[Y_{i+1} \middle| X_i\right] \mathbb{P}\left[X_{i+1} \middle| X_i\right], \tag{2.1}$$

for $i \geq 0$. In general, (2.1) does not have a closed form and must be propagated by a numerical approximation method. Moreover, (2.1) relies on knowing X_i, which is estimated from X_{i-1} by (2.1), which, in turn, relies on an estimate of X_{i-2} by (2.1). Hence, estimating X_{i+1} requires a numerical method that not only approximates (2.1), but also propagates the uncertainty in the estimates from $X_i, X_{i-1}, \ldots, X_0$. When such an estimation is performed, the estimation procedure is called a statisical filter. Popular statistical filters for this case include the extended Kalman Filter, the unscented Kalman Filter, the ensemble Kalman Filter and the Particle Filter, which are discussed in detail in [10].

Under certain circumstances, (2.1) can be propagated in closed form, which gives rise to the Kalman Filter. To describe the Kalman Filter, consider the situation where the transitions between hidden states are defined by $X_{i+1} = f(X_i, i) + \epsilon_i$, where $f : \mathbb{R}^h \times \mathbb{N} \cup \{0\} \to \mathbb{R}^h$; and $\epsilon_i \sim \mathcal{N}(0, \Sigma_i)$ are non-degenerate, independent random variables for all $i \geq 0$. Moreover, consider the observations defined by $Y_i = LX_i + \eta_i$, where $L \in \mathbb{R}^{h \times b}$; and $\eta_i \sim \mathcal{N}(0, \Gamma_i)$ are non-degenerate, independent random variables for all $i > 0$. Then, using Bayes' rule, we have that $\mathbb{P}\left[X_{i+1} \middle| X_i, Y_{i+1}\right]$ is

$$\mathcal{N}\left(f(X_i, i) + \Sigma_i L' \left(\Gamma_{i+1} + L\Sigma_i L'\right)^{-1} \left[Y_{i+1} - Lf(X_i, i)\right], \left[L'\Gamma_{i+1}^{-1} L + \Sigma_i^{-1}\right]^{-1}\right). \tag{2.2}$$

Thus, we have a description for X_{i+1} that involves another quantity, X_i.

However, as mentioned above, we must also propagate the uncertainty of the estimate of X_i into X_{i+1}. Unfortunately, this cannot be done in closed form unless $f(X_i, i)$ is a linear function of X_i; in which case, the uncertainty can be propagated in closed form, and $\mathbb{P}\left[X_{i+1} \middle| Y_{i+1}, Y_i, \ldots, Y_1, X_0\right]$ has the form of (2.2) but with Σ_i replaced by $\tilde{\Sigma}_i$ that contains the propagated uncertainty of the estimate of X_i. This estimation procedure is called the Kalman Filter.

The Kalman Filter has several glaring challenges. First, the Kalman Filter requires that X_0 is known precisely, or, more generally, that the mean of X_0 and its

covariance are known when X_0 is a random variable. Fortunately, the Kalman Filter is robust to this lack of precision [6]. Second, the Kalman Filter requires that the underlying relationship between X_i and X_{i+1} (i.e., $f(X_i, i)$) is linear, which is a highly specialized example. However, the Kalman Filter can be "extended" to the case of nonlinearity by using the Jacobian of $f(x, i)$ evaluated at X_i in $\tilde{\Sigma}_i$ and ignoring the additional variance added by the Jacobian. In general, this (Extended) Kalman Filter, has been extremely successful in practice [10].

Importantly, the Kalman Filter's success is not limited to replacing nonlinear relationships with an approximate linear relationship, but also to a majority of cases with systematic errors in the relationships between the hidden states, and to a majority of cases with the inclusion of deterministic chaos in the hidden state dynamics [10]. Intuitively, the Kalman Filter's ability to navigate these errors and approximations in the systematic error comes down to reinterpreting the Kalman Filter's estimate of X_{i+1} as a compromise between the approximate dynamics, which serve as a regularizing force in the estimation, and the observations, which serve as a corrective force in the estimation. Specifically, we can restate the Kalman Filter estimate, $\hat{X}_{i+1}$, as a proximal operator,

$$\hat{X}_{i+1} = \text{argmin}_z \left\{ \|Y_{i+1} - LZ\|^2_{\Gamma^{-1}_{i+1}} + \|Z - f(X_i, i)\|^2_{\tilde{\Sigma}^{-1}_i} \right\}. \tag{2.3}$$

From (2.3), the Kalman Filter is clearly balancing the information from the observation with information from the dynamics by their relative variances, Γ_i and $\tilde{\Sigma}_i$. Therefore, the Kalman Filter can reduce emphasis on inaccuracies in f by inflating its variance and giving more responsibility to Y_{i+1} in determining X_{i+1}. This interpretation of the Kalman Filter will be essential to our optimization application.

3. The Statistical-Filtering-Optimization Framework. Recall, in the previous section, we reviewed HMMs and Statistical filters in finite-dimensional vector spaces. In this section, we extend the HMMs and Statistical filters to infinite dimensional vector spaces for the purposes of optimization. As a stepping stone, we first describe HMMs and Statistical filters for functions along an iterate sequence.

3.1. Statistical Filtering of Functions along an Iterate Sequence. We begin by carefully developing the HMM that we will use to solve (1.1), and then we will develop the statistical filtering procedure for estimating this particular HMM. We conclude with several stability and convergence results.

3.1.1. The Hidden Markov Model. To define the states, let $\{x_0, x_1, \ldots\} \subset \mathbb{R}^d$. We define the states by

$$\{\mathbb{E}\left[f(x_i, W)\right], \mathbb{E}\left[\nabla f(x_i, W)\right] : i = 0, 1, \ldots\}, \tag{3.1}$$

where the gradients are evaluated with respect to the argument x. Moreover, we define the relationship between states by the approximation

$$\begin{bmatrix} \mathbb{E}\left[f(x_{i+1}, W)\right] \\ \mathbb{E}\left[\nabla f(x_{i+1}, W)\right] \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[f(x_i, W)\right] \\ \mathbb{E}\left[\nabla f(x_i, W)\right] \end{bmatrix} + \begin{bmatrix} \mathbb{E}\left[\nabla f(x_i, W)\right]' \\ 0 \end{bmatrix} (x_{i+1} - x_i) + \begin{bmatrix} \epsilon_i \\ \lambda_i \end{bmatrix}, \tag{3.2}$$

where the dynamics for the objective function are given by Taylor's theorem; and ϵ_i and λ_i are terms representing the systematic errors incurred by the approximation. We make two important remarks regarding (3.2). First, we can readily integrate higher-order derivatives into (3.2). Second, we can easily bound the approximation errors, ϵ_i and λ_i, as we now state.

LEMMA 3.1. *Suppose that* $\nabla \mathbb{E}\left[f(x,W)\right]$ *is* L*-Lipschitz continuous. Then,* $|\epsilon_i| \leq \frac{1}{2}L\left\|x_{i+1}-x_i\right\|_2^2$ *and* $\left\|\lambda_i\right\|_2 \leq L\left\|x_{i+1}-x_i\right\|_2$.

For the observations and their relationship to the states, we observe the concatenation of $f(x_i,W)$ and $\nabla f(x_i,W)$, which we assume are unbiased estimates of the objective function and its gradient. Moreover, we assume that the joint variance of these observations is given by $\Gamma(x_i) \succ 0$. Note, we will sometimes write $\Gamma_i = \Gamma(x_i)$. To reiterate, (3.1), (3.2) and the observations define a HMM for the objective function and its gradient *along a specific sequence of iterates* with approximate dynamics.

3.1.2. The Statistical Filter. We can now apply the (extended) Kalman Filter to estimate our HMM. The filter has three components that were implicitly described in (2.3) that we now define separately: the initial state estimate, $\{\hat{F}_0(x_0), \hat{G}_0(x_0)\}$, and its covariance, Σ_0; the analysis states, $\{\hat{F}_i^a(x_i), \hat{G}_i^a(x_i) : i \in \mathbb{N}\}$, and their covariances, $\{\bar{\Sigma}_i : i \in \mathbb{N}\}$; and the filtered states, $\{\hat{F}_i(x_i), \hat{G}_i(x_i) : i \in \mathbb{N}\}$, and their covariances, $\{\Sigma_i : i \in \mathbb{N}\}$.

Let $\{W_i : i = 0, 1, \ldots\}$ be independent random variables with the distribution of W. We now define the initial state as $\hat{F}_0(x_0) = f(x_0, W_0)$ and $\hat{G}_0(x_0) = \nabla f(x_0, W_0)$, from which it follows that the initial state has variance $\Sigma_0 = \Gamma_0$. From (3.2), we define the analysis states by

$$\begin{bmatrix} \hat{F}_i^a(x_i) \\ \hat{G}_i^a(x_i) \end{bmatrix} = \begin{bmatrix} \hat{F}_{i-1}(x_{i-1}) \\ \hat{G}_{i-1}(x_{i-1}) \end{bmatrix} + \begin{bmatrix} \hat{G}_{i-1}(x_{i-1})' \\ 0 \end{bmatrix} (x_i - x_{i-1}). \tag{3.3}$$

The covariance for the analysis states are computed in the following, straightforward result.

PROPOSITION 3.2. *Given the covariance of* $(\hat{F}_i(x_i), \hat{G}_i(x_i))$, Σ_i, *the covariance for* $(\hat{F}_{i+1}^a(x_{i+1}), \hat{G}_{i+1}^a(x_{i+1}))$ *is*

$$\bar{\Sigma}_{i+1}(h) = \begin{bmatrix} 1 & h' \\ 0 & I \end{bmatrix} \Sigma_i \begin{bmatrix} 1 & 0 \\ h & I \end{bmatrix}, \tag{3.4}$$

where $h = x_{i+1} - x_i$. *Moreover, if the filtered state is normally distributed, then the analysis state is normally distributed.*

While Proposition 3.2 gives an exact form for the covariance of the analysis states, this is not how we will use it. Recall that (3.2) is approximate and, following from the discussion in section 2, we will need to modify $\tilde{\Sigma}_i$ to account for the approximate state relationship. Specifically, we add to the covariance in Proposition 3.2 by a positive definite matrix $T_i(h)$. We will refer to the inflated version as $\tilde{\Sigma}_{i+1}(h)$.

Finally, the filtered state follows from (2.2) with $L = I$ and Σ_{i-1} replaced by $\tilde{\Sigma}_i = \bar{\Sigma}_i(x_{i+1} - x_i)$, which is given by

$$\begin{bmatrix} \hat{F}_i(x_i) \\ \hat{G}_i(x_i) \end{bmatrix} = \operatorname*{argmin}_{z \in \mathbb{R}^{d+1}} \left\{ \left\| z - \begin{bmatrix} f(x_i, W_i) \\ \nabla f(x_i, W_i) \end{bmatrix} \right\|_{\Gamma_i^{-1}}^2 + \left\| z - \begin{bmatrix} \hat{F}_i^a(x_i) \\ \hat{G}_i^a(x_i) \end{bmatrix} \right\|_{\tilde{\Sigma}_i^{-1}}^2 \right\}. \tag{3.5}$$

PROPOSITION 3.3. *Let* $h = x_i - x_{i-1}$. *Given the analysis state's covariance,* $\bar{\Sigma}_i = \bar{\Sigma}_i(h)$, *and denoting* $T_i = T_i(h) = \tilde{\Sigma}_i - \bar{\Sigma}_i$, *the covariance of the filtered state, defined by (3.5), is*

$$\left(\Gamma_i^{-1} + \tilde{\Sigma}_i^{-1}\right)^{-1} - \left(\Gamma_i^{-1} + \tilde{\Sigma}_i^{-1}\right)^{-1} \tilde{\Sigma}_i^{-1} T_i \tilde{\Sigma}_i^{-1} \left(\Gamma_i^{-1} + \tilde{\Sigma}_i^{-1}\right)^{-1}. \tag{3.6}$$

Moreover, if the analysis is normally distributed, then the filtered state is normally distributed.

However, in the optimization, we will see that such a procedure requires solving several d-dimensional linear systems in the line search. Therefore, if a trust region method is used or if the cost of generating a sample outweighs the cost of solving several d-dimensional linear systems, then this cost will be acceptable and (2.2) should be used. However, if a line search method is used or if the cost of generating samples is inexpensive to the cost of solving several d-dimensional systems, then we can use a marginal variant of (2.2) to compute $\hat{F}_i(x_i)$ and $\hat{G}_i(x_i)$ separately, but we will not give the details here.

3.1.3. Stability and Convergence of Estimates. The approximations used in defining the HMM and the statistical filter raise the question of if our estimates actually track the objective and gradient function. The following two results describe the behavior of the estimates within the relevant optimization context, that is, when $\{x_0, x_1, \ldots\}$ is a converging sequence. The first result states that when our approximation-motivated covariance correction term, $T_i(h)$ is $c \|h\|_2 I$ for some $c > 0$, we can guarantee that the variance of our estimates decays to zero. The second result states that when our covariance correction term is cI for $c > 0$, then we can guarantee that the bias of our estimates decays to zero. Together, these two results suggest that there is some limiting bias-variance trade-off that prevents the convergence of our estimators. We aim to address this concern in future efforts.

THEOREM 3.4. *Let $\{x_0, x_1, \ldots\} \subset \mathbb{R}^d$ be a sequence converging to x^*. Moreover, suppose that $\exists \gamma \in (0, \infty)$ such that $\sup\{\|\Gamma(x_i)\|_2\} < \gamma$. Now, let the analysis states be given by (3.3) and filtered states be given by (3.5), where $\tilde{\Sigma}_i(h) - \bar{\Sigma}_i(h) = c \|h\|_2 I$ for some $c > 0$. Then, $\lim \Sigma_i \to 0$.*

Proof. First, with $h_i = x_{i+1} - x_i$, note that $\|\tilde{\Sigma}_{i+1}\|_2 \leq \|\Sigma_i\|_2 (1 + \|h_i\|_2 + \|h_i\|_2^2) + c \|h_i\|_2$. Second, for symmetric, invertible matrices A and B, and for any $a \geq \|A\|_2$ and $b \geq \|B\|_2$, $\left(A^{-1} + B^{-1}\right)^{-1} \preceq \frac{ab}{a+b} I$. Therefore, by this preceding bound and by Propositions 3.2 and 3.3,

$$\|\Sigma_{i+1}\|_2 \leq \frac{\left[\|\Sigma_i\|_2 \left(1 + \|h_i\|_2 + \|h_i\|_2^2\right) + c \|h_i\|_2\right] \gamma}{\left[\|\Sigma_i\|_2 \left(1 + \|h_i\|_2 + \|h_i\|_2^2\right) + c \|h_i\|_2\right] + \gamma} \tag{3.7}$$

Fourth, for $\delta_i \in [0, \|\Sigma_i\|_2)$, if $\|h_i\|_2$ satisfies

$$\|\Sigma_i\|_2 [\gamma + \delta_i] \|h_i\|_2^2 + [\gamma + \delta_i] [c + \|\Sigma_i\|_2] \|h_i\|_2 + \|\Sigma_i\| [\delta_i - \|\Sigma_i\|_2] \leq 0, \tag{3.8}$$

then $\|\Sigma_{i+1}\|_2 \leq \|\Sigma_i\|_2 \frac{\gamma}{\delta_i + \gamma} \leq \|\Sigma_i\|_2$. For a contradiction, suppose that $\{\|\Sigma_i\|_2\}$ does not converge to 0. Then, there is an $\epsilon > 0$ such that for a subsequence $\{i_k\}$, $\|\Sigma_{i_k}\|_2 > 2\epsilon$. Then, with $\delta_{i_k} = \epsilon$, if $\|h_{i_k}\|_2$ satisfies

$$\gamma [\gamma + \epsilon] \|h_{i_k}\|_2^2 + (\gamma + c)(\gamma + \epsilon) \|h_{i_k}\|_2 - 2\epsilon^2 \leq 0, \tag{3.9}$$

then $\|\Sigma_{i_k}\|_2 \leq \|\Sigma_{i_k - 1}\|_2 \frac{\gamma}{\epsilon + \gamma} < \|\Sigma_{i_k - 1}\|_2$. Hence, for sufficiently large j, if $i \geq j$ then $\|\Sigma_i\|_2 > 2\epsilon$. However, for $i \geq j$, $\|\Sigma_i\|_2 \leq \|\Sigma_j\|_2 \left(\frac{\gamma}{\epsilon + \gamma}\right)^{i-j}$, which will be smaller than 2ϵ for $i - j$ large enough. This is our contradiction, and so $\|\Sigma_i\|_2 \to 0$. ∎

THEOREM 3.5. *Let $\{x_0, x_1, \ldots\} \subset \mathbb{R}^d$ be a sequence converging to x^*. Moreover, suppose that $\exists \gamma \in (0, \infty)$ such that $\sup\{\|\Gamma(x_i)\|_2\} < \gamma$. Now, let the analysis states be given by (3.3) and filtered states be given by (3.5), where $\tilde{\Sigma}_i(h) - \bar{\Sigma}_i(h) = cI$ for some $c > 0$. If $\nabla \mathbb{E}[f(\cdot, W)]$ is Lipschitz continuous then the bias of $\hat{F}_j(x_j)$ and $\hat{G}_j(x_j)$ converge to zero.*

Proof. First, let B_i denote the Euclidean norm of the bias of $(\hat{F}_i(x_i), \hat{G}_i(x_i))$. Then, using Lemma 3.1,

$$B_i \leq \left\|\Gamma_i(\Gamma_i + \tilde{\Sigma}_i)^{-1}\right\|_2 \left[(1 + \|h_i\|_2)B_{i-1} + \frac{1}{2}L\|h_i\|_2^2 + L\|h_i\|_2\right]. \tag{3.10}$$

Second, recall that if A and B are symmetric and, for $a > 0$, $A \succeq aI$ and $b \geq \|B\|_2$, then $A(A+B)^{-1} \preceq \frac{b}{b+a}I$. Therefore, since $\tilde{\Sigma}_i \succeq cI$,

$$B_i \leq \frac{\gamma(1 + \|h_i\|_2)}{\gamma + c}B_{i-1} + \frac{\gamma}{\gamma + c}\left[\frac{1}{2}L\|h_i\|_2^2 + L\|h_i\|_2\right]. \tag{3.11}$$

Finally, since $\|h_i\|_2 \to 0$, we have that $B_i \to 0$. ∎

3.2. Statistical Filtering over Function Spaces and Optimization. With the developments of the previous subsection, we are now ready to state our novel methodology for solving (1.1) that avoids the disadvantages of current techniques and retains their advantage. We will state this methodology in two steps.

First, we state how to extend the statistical filter to function spaces. Suppose we have $(\hat{F}_{i-1}(x_{i-1}), \hat{G}_{i-1}(x_{i-1}))$ with variance Σ_{i-1}. Then, we define function $\hat{F}_i(x)$ and $\hat{G}_i(x)$ for each x by

$$\begin{bmatrix}\hat{F}_i(x) \\ \hat{G}_i(x)\end{bmatrix} = \underset{z \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \left\{ \left\| z - \begin{bmatrix} f(x, W_i) \\ \nabla f(x, W_i)\end{bmatrix}\right\|_{\Gamma(x)^{-1}}^2 + \left\| z - \begin{bmatrix}\hat{F}_i^a(x) \\ \hat{G}_i^a(x)\end{bmatrix}\right\|_{\tilde{\Sigma}(x - x_{i-1})^{-1}}^2 \right\}. \tag{3.12}$$

By Lemma 3.1, (3.12) is only useful locally about x_{i-1}. However, this is exactly what we need because we will use an iterative optimization procedure, which is inherently local, to define x_i.

That is, for the second step, we define x_i to be an exact or approximate solution of

$$\begin{aligned} &\min_x \ \hat{F}_i(x) \\ &\text{subject to: } e_1'\Sigma_i(x)e_1 \leq \gamma e_1'\Sigma_i(x_{i-1})e_1, \end{aligned} \tag{3.13}$$

where $e_1 \in \mathbb{R}^{p+1}$ is the basis vector that has a one in its first component; and $\gamma > 1$. Effectively, the constraint requires that the variance of the objective does not grow too severely. Indeed, by the definition of $\tilde{\Sigma}_i(x)$, we can guarantee that there is a neighborhood of x_{i-1} in which this condition is guaranteed to hold. Therefore, we are guaranteed that a standard optimization solver can be used to solve (3.13) either completely or inexactly.

Once we have generated x_i, we can now repeat the procedure with $\hat{F}_i(x_i)$ and $\hat{G}_i(x_i)$ to, first, determine $\hat{F}_{i+1}(x)$ and $\hat{G}_{i+1}(x)$ using the statistical filter, and, second, to compute x_{i+1} using a standard deterministic optimizer. Thus, we have described a complete optimization methodology for addressing (1.1).

4. Numerical Experiments. Consider a study of 250 patients with the same disease who are each treated by one of 91 different physicians. In this study, we record the patient's state after treatment (not cured or cured); gender, g, (male or female); disposition prior to treatment, d, (not optimistic or optimistic); level of exercise, e (no exercise or does exercise); and the treating physician. Based on the outcomes of our study, we want to state a model that describes the probability of being cured based on the patient's disposition and level of exercise, while accounting for the possible variations between physicians (not just our 91 physicians). In this case, we might choose to model the probability of being cured as

$$p_{cured} = \frac{1}{1 + \exp\left[\beta_0 + \beta_d d + \beta_e e + \beta_g g + \rho\right]},$$

where β_0, β_d, β_e and β_g are unknown coefficients; d is one if the patient's disposition is optimistic and zero otherwise; e is one if the patient exercises at all and is zero otherwise; g is one if the patient is a female and zero otherwise; and ρ is a mean-zero, normally distributed random variable with an unknown variance σ_ρ^2 that represents the random deviation from the systematic behavior (described by the intercept, patient's behavior and patient's exercise level) based on the physician (see [1] for short overview of such models).

Let $p_{cured}^{(i)}$ denote the probability of patient i being cured as described by the model and let $y^{(i)}$ denote the state of the patient at the end of the treatment; specifically $y^{(i)} = 1$ if the patient is cured and is zero otherwise. Then, the likelihood function is

$$\mathcal{L}(\beta_0, \beta_d, \beta_e, \sigma_\rho) = \mathbb{E}\left[\prod_{i=1}^{250} (p_{cured}^{(i)})^{y^{(i)}} (1 - p_{cured}^{(i)})^{1-y^{(i)}}\right]. \tag{4.1}$$

We compute our estimates by minimizing $-\mathcal{L}(\beta_0, \beta_d, \beta_d, \sigma_\rho)$, which, in our notation is an example of (1.1).

Techniques for maximizing (4.1) are of two varieties: a finite approximation to the integral, or an approximation of the integrand that results in an analytic form of the objective function [11]. Unfortunately, these techniques are known to perform poorly when the dimension of the random components increase, when a normal approximation to the integrand is invalid, and when using a large number of groups with few measurements per group (such as in our example) [11].

For our experiment, we simulate this study and compare three techniques for solving (4.1). We use the de facto standard solver implemented in the R Programming Language in the package `lme4` [1]. Based on our approach, we use a Kalman Filter Gradient Descent with Line Search (KF-GD-LS) with $T(h) = c \|h\|_2 I$ with $c > 0$, and we use a Kalman Filter Gradient Descent with Trust Region (KF-GD-TR) with $T(h) = cI$ with $c > 0$. The observations for our solver are generated by a Monte Carlo approximation to the integral in (4.1) with 100 random samples. In order to compare the techniques to some "truth" we compute (4.1) and its derivative using a Monte Carlo approximation with 10,000 random samples, which has a variance on the order of 10^{-4}.

The comparison of our solvers to the "truth" and to the `lme4` solution are presented in Figures 4.1 and 4.2. There are several notable features. First, the KF-GD-LS approach outperforms, and the KF-GD-TR approach performs at least as well as the `lme4` solver, which is shown is red. Second, as stated in Theorem 3.4, the variance of the KF-GD-LS approach is converging to zero, as indicated by the thinning of the blue

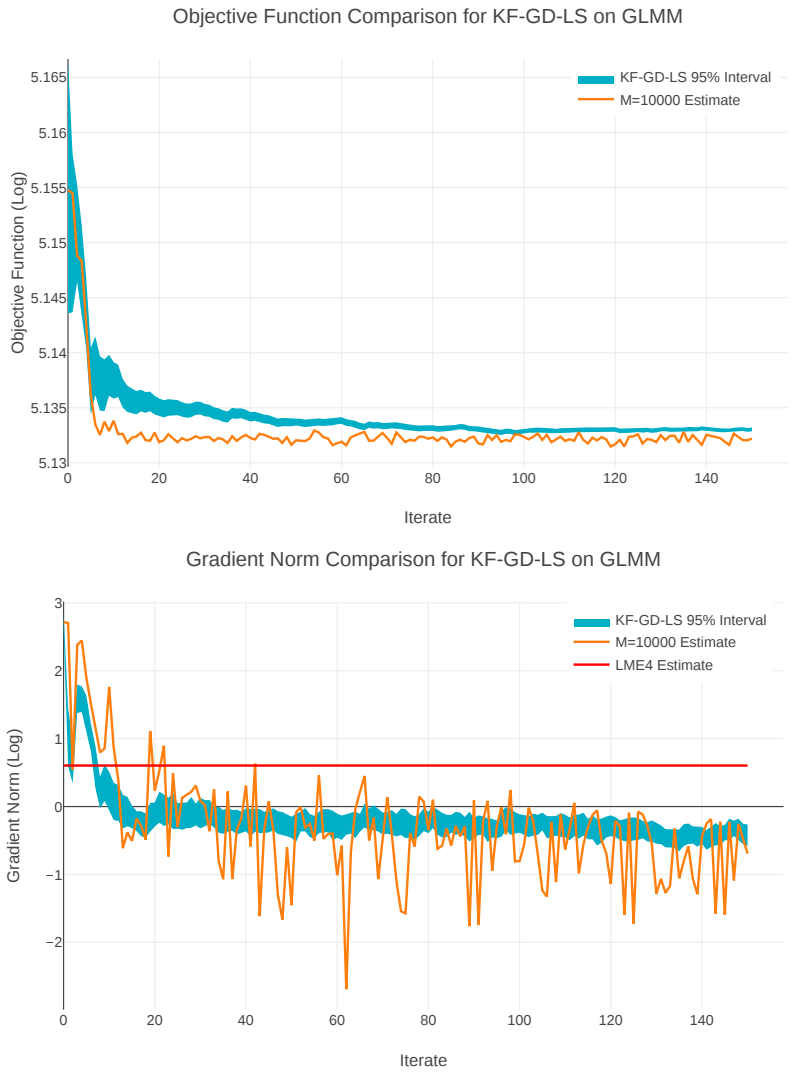

FIG. 4.1. *A comparison of the de facto solver and our KF-GD-LS approach, for which* $T(h) = c\,\|h\|_2\, I$. *A 95% confidence interval is computed for the KF-GD-LS approach based on its covariance estimates.*

line, which represents a 95% confidence interval computed from the covariance matrix estimates. However, as we discussed, the bias of the KF-GD-LS approach does not decay, which is particularly obvious in the gradient norm plot in Figure 4.1. Similarly, as stated in Theorem 3.5, we do not see a severe bias problem for the KF-GD-TR approach, but we also see that the variance of the objective and gradient estimates do not decay.

5. Conclusions. Motivated by the shortcomings of existing methodologies for solving (1.1), in this work, we introduced a novel framework that combines statistical filters and deterministic optimization methods to generate solvers for (1.1). After detailing this framework, we provided initial theoretical results regarding the bias and

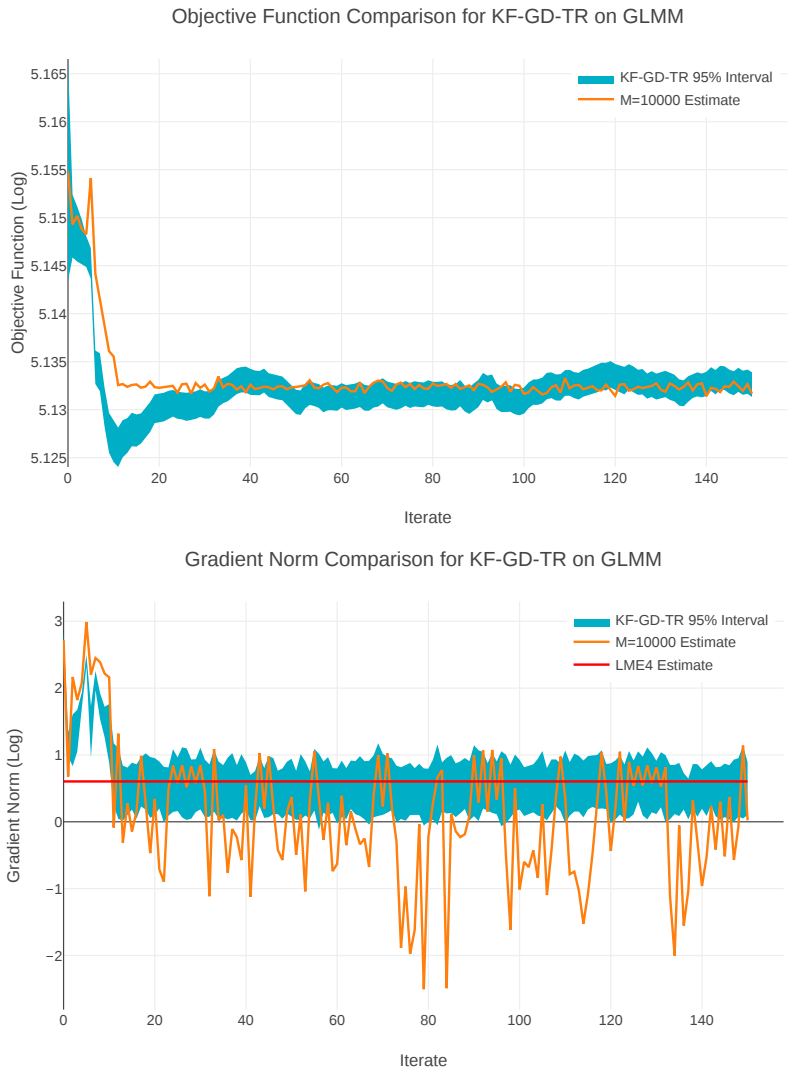

FIG. 4.2. *A comparison of the de facto solver and our KF-GD-LS approach, for which* $T(h) = c\,\|h\|_2\, I$. *A* 95% *confidence interval is computed for the KF-GD-TR approach based on its covariance estimates.*

variance of the statistical filtering estimates for the objective and gradient function. In particular, we showed that either the bias or variance will decay depending on the form of the covariance correction term, which was reinforced by the numerical experiments. In the future, our main effort will be to improve the estimation methodology so that both the bias and the variance are guaranteed to decay to zero, which implies convergence of the estimators in probability and L^2.

Moreover, we used the framework to generate two different optimization methodologies for inverting the generalized linear mixed effect model. We compared the methodologies to the de facto standard used in R for performing the same task, and we observed that our methodologies either outperform or perform just as well as the de facto standard.

In the future, we will aim to fortify and extend this methodology in several ways. From the filtering aspect, as mentioned, we will further develop the estimation methodology to ensure that both the bias and the variance are guaranteed to decay to zero; we will experiment with other local representations and filtering schemes to provide users with a detailed understanding of the benefits and disadvantages of using different filters within our framework; finally, while Hessian information is easy to integrate, we will study how using quasi-Newton Hessian estimates impacts the overall theoretical and convergence properties of our methodology. From the optimization perspective, we will begin studying this framework for more complex optimization problems such as those problems with convex or nonlinear constraints and those problems that do not have gradient information.

REFERENCES

[1] D. Bates, M. Mächler, B. Bolker, and S. Walker, *Fitting linear mixed-effects models using lme4*, Journal of Statistical Software, 67 (2015), pp. 1–48, https://doi.org/10.18637/jss.v067.i01.

[2] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, arXiv preprint arXiv:1606.04838, (2016).

[3] E. Brochu, V. M. Cora, and N. De Freitas, *A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning*, arXiv preprint arXiv:1012.2599, (2010).

[4] D. P. Kouri and T. M. Surowiec, *Risk-averse pde-constrained optimization using the conditional value-at-risk*, SIAM Journal on Optimization, 26 (2016), pp. 365–396.

[5] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on optimization, 19 (2009), pp. 1574–1609.

[6] V. Patel, *Kalman-based stochastic gradient method with stop condition and insensitivity to conditioning*, SIAM Journal on Optimization, 26 (2016), pp. 2620–2648.

[7] V. Patel, *The impact of local geometry and batch size on the convergence and divergence of stochastic gradient descent*, arXiv preprint arXiv:1709.04718, (2017).

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

[9] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications: with R examples*, Springer Science & Business Media, 2006.

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

[11] F. Tuerlinckx, F. Rijmen, G. Verbeke, and P. Boeck, *Statistical inference in generalized linear mixed models: A review*, British Journal of Mathematical and Statistical Psychology, 59 (2006), pp. 225–255.