Mathematical Biosciences and Engineering

https://www.aimspress.com/journal/mbe

MBE, 23(7): 2018–2054.
DOI: 10.3934/mbe.2026074
Received: 20 November 2025
Revised: 19 January 2026
Accepted: 18 March 2026
Published: 09 July 2026

Research article

Novel optimization techniques for inferring heterogeneous population dynamics

Chenyu Wu[1,*], **Nuozhou Wang**[1], **Casey Garner**[2], **Kevin Leder**[1] **and Shuzhong Zhang**[1]

[1] Department of Industrial & Systems Engineering, University of Minnesota, USA

[2] School of Mathematics, University of Minnesota, USA

* **Correspondence:** Email: wu000766@umn.edu; Tel: +17188665568.

Abstract: In this paper, we introduce a new optimization algorithm that is well suited to solve parameter estimation problems that arise when inferring heterogeneous population dynamics. In these estimation problems, parameter estimation is complicated by the presence of two types of constraints: inequality constraints (e.g., non-negativity and boundedness of rates (so-called box-constraints)) and equality constraints that arise due to the need of the population fractions to sum to one. We call our new method cubic regularized Newton with affine scaling (CRNAS). In contrast to so-called first-order methods, which solely rely on the gradient of the objective function, our method utilizes the Hessian of the objective. As a result, it is able to focus on points that satisfy the second-order optimality conditions, as opposed to first-order methods that simply converge to critical points. This is an important feature in parameter estimation problems, where the objective function is often non-convex; as a result, there can be many critical points, which makes it nearly impossible to identify the global minimum. We use an affine scaling approach to handle a wide class of constraints, including equality constraints. We establish that CRNAS identifies a point that satisfies ϵ-approximate second-order optimality conditions within $O(\epsilon^{-3/2})$ iterations. Finally, we compare CRNAS with MATLAB's optimization solver *fmincon* on three different test problems. These test problems all feature mixtures of heterogeneous populations, a problem setting that CRNAS is particularly well-suited for. Our numerical simulations show that CRNAS has a favorable performance, thereby performing comparable, if not better than, *fmincon* in accuracy and computational cost for most of our examples.

Keywords: nonlinear programming; nonconvex optimization; parameter estimation; heterogeneous population dynamics; branching process

1. Introduction

Mathematical modeling plays a critical role in understanding and quantifying biological systems. In particular, it enables the quantification of dynamics, thus providing the ability to predict potential outcomes and guide future studies. An essential component of successful mathematical modeling is parameter estimation. Here, the goal is to train and/or fit a given mathematical model to observed data. The most common method of parameter estimation follows the maximum likelihood framework, where one formulates a statistical model with associated parameters and maximizes a likelihood function over the parameter space to fit the model to provided data. Thus, parameter estimation problems often reduce to optimization problems.

Parameters in mathematical models for biological systems usually have a physical interpretation. As a result, the values they can assume are often restricted based on our knowledge of the natural world. For example, non-negativity and upper-bound constraints are often desired, if not necessary. This is in contrast to neural network models, where the parameters are often unconstrained. The presence of constraints on the parameter space adds non-trivial difficulties to the derived parameter estimation optimization problem. For instance, for constrained problems, one cannot simply apply gradient descent since the updates could take the parameters out of their feasible region.

A common feature of mathematical models that arise in the study of biological systems is their highly nonlinear and non-convex dependence on parameters [1]. For example, the parameter sensitivities in a system of nonlinear differential equations also satisfy a nonlinear system of differential equations [2]. Even when a mathematical model is available in an explicit form, it often exhibits nonlinear dependence on its parameters. For example, the logistic growth model, $y(t) = L/(1 + e^{-k(t-t_0)})$, nonlinearly depends on the parameters k and t_0. Such nonlinear dependence on the parameters may sometimes cause significant challenges, as the objective function can become highly sensitive or even undefined when the parameters approach or violate the constraint boundaries. Therefore, optimization algorithms such as the active-set method [3] or the Augmented Lagrangian method [4], which may stop at or violate the constraints, are not ideal to solve these problems. An ideal optimization algorithm should always search within the feasible region.

Another phenomenon to consider when modeling biological systems is population heterogeneity. For example, when studying how cells respond *in vitro* to chemotherapy, a heterogeneous response is often observed [5]. To model this type of heterogeneity, one often assumes there is a known number of distinct subpopulations, S, present at unknown proportions, $\{p_i\}_{i=1}^{S}$, each with distinct characteristics. The presence of these unknown proportions introduces a new challenge to the parameter estimation problem, namely equality constraints. In particular, we know *a priori* that any valid set of proportions must satisfy the following conditions: $\sum_{i=1}^{S} p_i = 1$ and $p_i > 0$ for $i = 1, \ldots, S$. Such constraints are typically expressed as linear constraints of the form $Ax = b$, which add nontrivial difficulties in non-convex optimization [6], particularly when higher-order derivatives of the objective function are incorporated into the algorithm design.

Taken together, these considerations show that parameter estimation in biological models using a maximum likelihood framework reduces to a nonlinear, non-convex, constrained optimization problem [7–9]. The loss of convexity presents a significant challenge. Convexity in optimization enables us to infer global properties from local properties, (i.e., local minimums are also global minimums). This ability to infer global properties from local ones is lost in the absence of convexity.

Global optimal solutions are preferred, of course, and this is no different in parameter estimation. Multiple global optimization frameworks have been suggested to address this issue [10, 11]. These frameworks propose strategies to locate the global optimum, while relying on local optimization algorithms that typically converge to locally optimal solutions. A frequently employed technique is the multi-start approach; this technique suggests to solve the optimization problem starting from randomly selected initial points and obtaining multiple local solutions. Among these local solutions, it can be argued that the best one corresponds to the global optimum—provided the problem only admits a finite number of local optima. However, verifying the finiteness of local solutions is particularly challenging in nonlinear, non-convex, constrained, continuous optimization problems. In summary, modern global optimization frameworks heavily rely on computing multiple local solutions of the constrained optimization problem, thus emphasizing the need for efficient and accurate local methods.

Many of the likelihood functions utilized in the study of mathematical biology have accessible higher-order derivative information [12, 13]; however, numerous optimization algorithms do not take advantage of this. Procedures such as gradient descent and quasi-Newton methods only use the gradient information of the objective function, which categorizes them as first-order methods. For this reason, these approaches typically only yield solutions that satisfy the first-order optimality conditions, which might not be locally optimal. For instance, when minimizing the simple function $f(\theta) := \sin\left(\frac{\pi}{2}\theta\right)$ subject to the constraint $-1 \leq \theta \leq 1$, first order stationary points (FOSPs) that satisfy the first-order optimality conditions are not necessarily local minima. In contrast, second-order stationary points (SOSPs) that satisfy the second-order optimality condition are preferable, as they filter out some of the non-optimal solutions. In the parameter estimation problems we consider, it is possible to compute the Hessian of the objective function and utilize this information to obtain better solutions. Therefore, in this work, we explore the advantages of optimization algorithms utilizing second-order information and develop a new algorithm. Our proposed method is called cubic regularized Newton based on affine scaling (CRNAS). This method combines the ideas of the cubic regularized Newton's method and affine scaling in order to provide a novel second-order scheme capable of solving constrained optimization problems. An important advantage of CRNAS over first-order methods is that it finds solutions that satisfy second-order optimality conditions. Due to the high levels of nonlinearity in mathematical biology models, many solutions that satisfy first-order optimality conditions may be suboptimal; therefore, avoiding convergence to such points is critical, and achieving second-order optimality can significantly improve the solution quality [4, 14, 15]. We provide a convergence analysis for CRNAS, thereby obtaining the best possible iteration complexity bound for the class of models we study. Numerical experiments show that a simple implementation of CRNAS is already competitive with MATLAB's state-of-the-art solvers. Another advantage of CRNAS for parameter estimation in heterogeneous populations is its ability to enforce the constraint feasibility. By maintaining feasibility throughout optimization, CRNAS avoids potential instabilities or ill-defined objective values near the boundary or outside the feasible region, and ensures that the final estimates remain physically meaningful.

The paper is structured as follows. Section 2 develops CRNAS for a general class of constrained non-convex problems. We discuss the two methodologies grafted together to construct CRNAS, the cubic regularized Newton's method and the affine scaling method, and present the convergence theory. We provide the proof for the main results in Appendix A.1. Sections 3.1 and 3.2 present

numerical experimentation with CRNAS based on two case studies: Section 3.1 investigates a cancer drug response estimation problem, and Section 3.2 studies an application in heterogeneous logistic growth estimation. The paper concludes in Section 4 with a discussion of the advantages and limitations of CRNAS, along with a proposed implementation scenario. The appendices include the technical details of our analysis, an exposition on CRNAS's implementation, and setup of our numerical experiments.

2. Materials and methods

2.1. Optimization framework

In a parameter estimation problem, one has a function that maps inputs and a parameter set to a desired output. Then, for a given dataset, one is interested in finding the best fitting (in some sense) parameter set. How one defines "best fitting" is a choice that the modeler makes based on their understanding of the dataset and its generation. In particular, assume we have a function f that takes the inputs x and parameter set θ and generates a prediction of a desired output, $y = f(x;\theta)$. If we have a dataset of paired inputs and outputs, $(x_i, y_i)_{i=1}^n$, and assume a simple statistical model that generates the outputs, $y_i = f(x_i;\theta) + Z_i$, where Z_i are independent and identically distributed random variables with probability density function ϕ that are used to represent the potential observation noise, then we can write the negative log-likelihood of the parameter set given by the observed data as follows:

$$L\left(\theta \mid (x_i, y_i)_{i=1}^n\right) = -\sum_{i=1}^{n} \log \phi\left(y_i - f(x_i;\theta)\right).$$

To complete the parameter estimation problem, we must minimize the function $L\left(\theta \mid (x_i, y_i)_{i=1}^n\right)$ over θ; however, as mentioned before, the components of θ might have a physical interpretation and therefore be constrained. For example, they might be constrained to be non-negative, bounded, or, as stated in the introduction, they might be proportions that sum to one. As a result, we reduced the parameter estimation problem to a constrained optimization problem. Note that when writing the negative log-likelihood, we will often drop the explicit dependence on the observed data.

In this work, we focus on creating algorithms to numerically solve a broad class of constrained optimization problems that arise from parameter estimation problems of the following form:

$$\begin{array}{ll} \min_{\theta\in\mathbb{R}^n} & L(\theta) \\ \text{s.t.} & A\theta = b, l \le \theta \le u, \end{array} \tag{2.1}$$

where the objective function $L : \mathbb{R}^n \to \mathbb{R}$ is possibly non-convex, and the domain is the intersection of a linear and box constraint. Observe, all the examples given for constraints on the parameters can be represented by the conditions in (2.1), and we can actually further generalize our optimization framework by noting that the box constraint, $l \le \theta \le u$, can be rewritten as a conic constraint. Letting $\theta_1 := \theta - l$ and $\theta_2 := u - \theta$, we see the conditions $\theta_1,\ \theta_2 \ge 0$ and $\theta_1 + \theta_2 = u - l$ are equivalent to the box constraint. Therefore, we can replace the box constraint with an additional linear equality constraint and restrictions to the non-negative orthant (i.e., $\mathbb{R}^n_+ := \{x = (x_1, \ldots, x_n) \mid x_i \ge 0,\ i = 1, \ldots, n\}$).

The set $\mathbb{R}^n_+$ is an example of a pointed convex cone, that is, a set $\mathcal{K} \subseteq \mathbb{R}^n$ such that for all $x, y \in \mathcal{K}$ and $t \ge 0$: $tx \in \mathcal{K}$, $x + y \in \mathcal{K}$, and $-\mathcal{K} \cap \mathcal{K} = \{0\}$. Additionally, if the span of $\mathcal{K}$ is the entire

ambient space, then we say the cone is solid. Thus, due to this equivalence, we shall instead develop our algorithm to solve the more general class of minimization problems:

$$\begin{array}{ll} \min_{\theta\in\mathbb{R}^n} & L(\theta) \\ \text{s.t.} & A\theta = b, \theta \in \mathcal{K}, \end{array} \tag{2.2}$$

where $\mathcal{K} \subseteq \mathbb{R}^n$ is a pointed convex solid cone, and $L : \mathcal{K} \to \mathbb{R}$ is smooth but possibly non-convex. For a first-order method, one typically expects to find a first-order solution, which is defined as $\nabla L(\theta) = 0$ without constraints. However, like the examples in Sections 3.1 and 3.2, functions may have many saddle points which satisfy first-order conditions while failing to be local minimums, and second-order methods are ideal to avoid converging to such saddle points. In the optimization literature, second order methods that converge to second-order stationary points have been developed to solve non-convex problems [16–18]; however, relatively few papers in the literature deal with developing second-order methods for non-convex constrained models. Independent of our work, Dvurechensky and Staudigl [19] proposed a similar approach to ours to solve non-convex constrained optimization models. The main difference between our work and theirs lies in the fact that we use the Dikin ellipsoid and a cubic regularization of the objective to define our subproblems, while they only used a cubic regularization of the objective to form their subproblems.

For the constrained optimization problems defined in Eq (2.2), defining the corresponding first-order (FOSP) and second-order stationary points (SOSP) is nontrivial. In this paper, we adopt the definitions proposed in [20]. Moreover, the authors of [20] introduced a so-called Newton-conjugate gradient (Newton-CG) based barrier method to find an $(\epsilon, \sqrt{\epsilon})$-SOSP for non-convex, conically constrained models with $O(\epsilon^{-3/2})$ Cholesky factorizations and $\tilde{O}(\epsilon^{-3/2}\min\{n, \epsilon^{-1/4}\})$ other fundamental operations. However, the method proposed in [20] is based on the central path-following approach, which requires a target point on the central-path to be pre-specified. The so-called central-path is a parameterized trajectory defined by minimizing the objective plus a varying positive parameter times the barrier function [21]. However, when solving the non-convex model, the above-mentioned function which defines the central-path is no longer convex; hence, its minimizer may not be unique or continuous. In this sense, the so-called "central-path" is no longer guaranteed to be well-defined for a nonconvex optimization model. On the other hand, the authors of [19] applied a cubic regularized model to get a $O(\epsilon^{-3/2})$ worst-case iteration bound. Additionally, CRNAS obtains a $O(\epsilon^{-3/2})$ worst-case iteration bound to ϵ-SOSP for (2.2); CRNAS is simple to implement and is well suited for the situation when the Hessian is easily obtainable. Additionally, our approach avoids the need to solve a possibly ill-conditioned Newton's equation; we only require the solution to a simple subproblem which we discuss in Appendix A.2. Next, we introduce the notion of self-concordant barrier functions for conic constraints, which play an important role in our analysis, and the corresponding notion of ϵ-FOSP and ϵ-SOSP which arise from them.

2.2. Logarithmic homogeneous self-concordant barrier functions

Logarithmic homogeneous self-concordant barrier functions [21] play an important role in interior-point methods. In this paper, we assume the cone $\mathcal{K}$ is convex pointed, solid, and has a logarithmic homogeneous self-concordant barrier function $B : \text{int}(\mathcal{K}) \to \mathbb{R}$, that is, B is convex, three-times continuously differentiable over $\text{int}(\mathcal{K})$, $B(\theta_k) \to \infty$ for all sequences $\{\theta_k\}_{k\in\mathbb{N}} \subseteq \text{int}(\mathcal{K})$ which converge to a point on the boundary of $\mathcal{K}$, and for any $\theta \in \text{int}(\mathcal{K})$, $\tau > 0$, and direction $h \in \mathbb{R}^n$, the following

properties are satisfied:

$$|\nabla^3 B(\theta)[h,h,h]| \leq 2(\nabla^2 B(\theta)[h,h])^{3/2},$$

$$B(\tau\theta) = B(\theta) - D\log\tau,$$

where D is a positive constant. As an example, $B(\theta) = -\sum_{i=1}^{n}\log(\theta_i)$ is a logarithmic homogeneous self-concordant barrier function for the cone $\mathbb{R}^n_+$ with $D = n$.

An important aspect of barrier functions comes from the fact they can define induced local norms. These norms are standard in the conic optimization literature and play a crucial role in our algorithm development and analysis. Following convention, we let

$$\begin{aligned} \|v\|_\theta &:= \left(v^\top \nabla^2 B(\theta) v\right)^{1/2} \\ \|v\|_\theta^* &:= \left(v^\top \left[\nabla^2 B(\theta)\right]^{-1} v\right)^{1/2} \\ \|C\|_\theta^* &:= \max_{\|v\|_\theta \leq 1} \|Cv\|_\theta^* \end{aligned}$$

for all $v \in \mathbb{R}^n$ and $C \in \mathbb{R}^{n\times n}$. Note, unless otherwise stated, $\|\cdot\|$ denotes the regular Euclidean norm. Following the pattern of [20], a feasible solution $\hat{\theta} \in \text{int}(\mathcal{K})$ with $A\hat{\theta} = b$ is said to be an ϵ-approximate first-order stationary point (ϵ-FOSP) for (2.2) if

$$\text{dist}\left((\nabla^2 B(\hat{\theta}))^{-1/2}\nabla L(\hat{\theta}), (\nabla^2 B(\hat{\theta}))^{-1/2}\text{Range}(A^\top)\right) = O(\epsilon),$$

which means that the angle between $(\nabla^2 B(\hat{\theta}))^{-1/2}\nabla L(\hat{\theta})$ and the orthogonal complement subspace of $(\nabla^2 B(\hat{\theta}))^{-1/2}\text{Range}(A^\top)$, namely $(\nabla^2 B(\hat{\theta}))^{1/2}\text{Null}(A)$, is of order ϵ. The latter statement can be restated as in terms of the local norms as

$$|(\nabla L(\hat{\theta}))^\top d| \leq \epsilon\|d\|_{\hat{\theta}} \tag{2.3}$$

for all $Ad = 0$. Moreover, a solution $\hat{\theta}$ is called an ϵ-approximate second-order stationary point (ϵ-SOSP) if in addition to being an ϵ-FOSP, the point $\hat{\theta}$ also satisfies

$$d^\top \nabla^2 L(\hat{\theta}) d \geq -\sqrt{\epsilon}\|d\|_{\hat{\theta}}^2$$

for all $Ad = 0$; see Remark 1 (ii) of [20] for further details. Next, we introduce the two methods which inspired our algorithm design.

2.3. Cubic regularized Newton's method

The cubic regularized Newton's method [22] was first proposed by Griewank to solve smooth unconstrained optimization problems [23]. Their approach adds a cubic regularization term to the second-order Taylor expansion of the objective function about the current iterate, θ^k, and solves this subproblem to compute the next iterate, i.e.,

$$\theta^{k+1} \in \arg\min_{\theta\in\mathbb{R}^n}\left(\left\langle \nabla L\left(\theta^k\right), \theta - \theta^k\right\rangle + \frac{1}{2}\nabla^2 L\left(\theta^k\right)[\theta - \theta^k]^2 + \frac{M}{6}\left\|\theta - \theta^k\right\|^3\right). \tag{2.4}$$

If the objective function has a gradient Lipschitz Hessian with constant $L_H \geq 0$ and $M \geq L_H$, then the above subproblem amounts to minimizing a cubic upper bound of the objective function about θ^k.

Hence, under mild assumptions, the cubic regularized Newton monotonically decreases the value of the objective function.

This method boasts multiple desirable qualities. The algorithm globally converges to second-order stationary points and computes a point that satisfies the approximate first-order unconstrained optimally conditions (i.e., $\|\nabla L(\theta)\| \leq \epsilon$, within $O(\epsilon^{-3/2})$ iterations), which bests gradient descent, and the method quadratically converges near strict local minimums [22]. Furthermore, though (2.4) is generally a non-convex problem, the subproblem is actually equivalent to minimizing a convex function in one variable. Thus, the cubic regularized Newton is an implementable, globally convergent, and locally fast converging method to second-order stationary points. For these reasons, we seek to incorporate aspects of this procedure into our design.

2.4. Affine scaling method

The affine scaling method [24–26] is an algorithm to solve linear programming problems by rescaling the variables and constraining the next iterate to lie within a ball contained inside the cone $\mathbb{R}^n_+$ to get a better iterate at a reduced computational cost. Consider the following linear program:

$$\begin{array}{ll} \min_{\theta\in\mathbb{R}^n} & c^\top\theta \\ \text{s.t.} & A\theta = b, \theta \geq 0. \end{array}$$

In each iteration, we scale the problem based on the current point $\theta^k > 0$. Let $\theta' = D_k^{-1}(\theta - \theta^k) = D_k^{-1}\theta - \mathbf{1}_n$, where $D_k = \text{Diag}(\theta^k)$ is a diagonal matrix with diagonal elements θ^k, and $\mathbf{1}_n$ is the vector of all ones in $\mathbb{R}^n$. Then, the non-negativity constraint is equivalent to $\theta' \geq -\mathbf{1}_n$. We desire to have the next iterate stay inside the interior of $\mathbb{R}^n_+$, so we replace $\theta' \geq -\mathbf{1}_n$ with a ball constraint, that is, we instead enforce $\|\theta'\| \leq 1 - \alpha$, where $0 < \alpha < 1$. Then, this ensures that the next iterate cannot lie on the boundary of the cone with the proximity to the boundary dictated by α (i.e., α close to zero can yield new iterates near the boundary). With this new constraint and change-of-variable, we form the following subproblem:

$$\begin{array}{ll} \min_{\theta\in\mathbb{R}^n} & (D_k c)^\top\theta' \\ \text{s.t.} & AD_k\theta' = 0, \|\theta'\| \leq 1 - \alpha. \end{array} \tag{2.5}$$

The main benefit of (2.5) is that it has a closed formed solution. Therefore, the affine scaling method proceeds by forming and solving (2.5) and using the variable transformation to obtain the next iterate. Utilizing the notation introduced in Section 2.2, then we can concisely write the affine scaling update as follows:

$$\begin{array}{rcll} \theta^{k+1} & = & \arg\min_{\theta\in\mathbb{R}^n} & c^\top\theta \\ & & \text{s.t.} & A\theta = b, \|\theta - \theta^k\|_{\theta^k} \leq 1 - \alpha \end{array} \tag{2.6}$$

where $B(\theta) = -\sum_{i=1}^n \log(\theta_i)$ is the barrier function used to define the norm in the constraint.

2.5. Cubic regularized Newton with affine scaling (CRNAS)

Our new method seeks to combine the ideas of the cubic regularized Newton and affine scaling to develop a procedure which solves (2.2). The crux of our approach is to utilize the affine scaling method to handle the constraints, thereby leveraging the concept of homogeneous self-concordant barrier functions to deal with the general conic constraints, but replace the linear objective in (2.6)

with the local cubic approximation of the objective function in (2.4). Thus, our method generates new iterates by solving the following subproblem:

$$\theta^{k+1} = \underset{A\theta=b,\|\theta-\theta^k\|_{\theta^k}\leq 1-\alpha}{\arg\min} \left(\left\langle \nabla L\left(\theta^k\right), \theta-\theta^k \right\rangle + \frac{1}{2}\nabla^2 L\left(\theta^k\right)[\theta-\theta^k]^2 + \frac{M}{6}\left\|\theta-\theta^k\right\|_{\theta^k}^3 \right), \tag{2.7}$$

where M is a positive number. Since our method brings together these two different procedures, we coined it CRNAS. A precise description of CRNAS is provided below.

Cubic regularized Newton with affine scaling (CRNAS)

Step 0: Provide an interior point θ^0 (i.e., $A\theta^0 = b$ and $\theta^0 \in \text{int}(\mathcal{K})$); choose the constants $\eta > 0$, $M > 0$, and $\alpha \in (0, 1)$; set $k = 0$

Step 1: Solve the following subproblem and proceed to Step 2:

$$\theta^{k+1} = \underset{\theta:A\theta=b,\|\theta-\theta^k\|_{\theta^k}\leq 1-\alpha}{\arg\min} \left(\left\langle \nabla L\left(\theta^k\right), \theta-\theta^k \right\rangle + \frac{1}{2}\nabla^2 L\left(\theta^k\right)[\theta-\theta^k]^2 + \frac{M}{6}\left\|\theta-\theta^k\right\|_{\theta^k}^3 \right)$$

Step 2: If $\|\theta^{k+1} - \theta^k\|_{\theta^k} < \eta$, then let $K = k + 1$ and stop; otherwise, go back to Step 1 with $k = k + 1$

Though (2.7) appears to be more difficult to solve than (2.4), this subproblem forming the backbone of CRNAS is solvable. A theoretical analysis and practical approach to solving (2.7) is detailed in Appendix A.2. In the next section, we provide an overview of the convergence guarantees we have for CRNAS; the technical proofs of the results are left to Appendix A.1.

2.6. Overview of the complexity analysis of CRNAS

First, we declare that CRNAS is well-defined in the sense all of the iterates produced by the algorithm remain inside the cone $\mathcal{K}$. The following lemma guarantees this follows from the constraints in the subproblem.

Lemma 2.1. *(Theorem 2.1.1, [21]) Let B be self-concordant on $\mathcal{K}$, and let $\theta_0 \in \text{int}(\mathcal{K})$, then, $\{\theta : \|\theta - \theta_0\|_{\theta_0} < 1\} \subset \text{int}(\mathcal{K})$.*

For our convergence theory to hold, we assume the following scaled Lipschitz smoothness of the second-order derivative of the objective.

Assumption 1. *There exists a constant $\beta \geq 0$ such that for all $x, y \in \text{int}(\mathcal{K})$*

$$\|\nabla^2 L(y) - \nabla^2 L(x)\|_x^* \leq \beta\|y - x\|_x.$$

We note that this assumption is not original to us but has precedence in the literature (e.g., [20]). As an example, quadratic functions can trivially satisfy the assumption, and the logarithmic homogeneous self-concordant barrier functions can locally meet the assumption. It has been pointed out [6] that checking the second-order stationarity without scaling is an NP-Hard problem. Therefore, this condition with scaling is reasonable. Under these limited conditions, we present our main complexity theorem; the proof can be found in Appendix A.1.

Theorem 2.2. *If Assumption 1 holds and CRNAS is applied to* (2.2) *with* $M = 2\beta$ *and* $\eta = \min\{1-\alpha, \epsilon^{1/2}\alpha^{1/2}M^{-1/2}, \frac{1}{\sqrt{2}}\epsilon^{1/2}\alpha^2M^{-1}\}$*, then the algorithm will stop within* $K \leq 12(L(\theta^0)-L^*)\eta^{-3}+1$ *iterations with* θ^K *at an* ϵ*-SOSP.*

Therefore, in view of the discussions on the optimality conditions in Subsection 2.2, Theorem 2.2 implies that in at most $O(\epsilon^{-3/2})$ iterations, CRNAS is guaranteed to find an ϵ-SOSP. Therefore, for the highly nonlinear and non-convex optimization problems coming from parameter estimation problems, CRNAS is able to avoid the many sub-optimal first-order stationary points which lead to a proposed solution of often superior quality compared to first-order methods.

It is worth remarking that one can devise a first-order version of CRNAS which uses a quadratic approximation of the objective function in the subproblem rather than a cubic approximation. A similar convergence result can be derived for this procedure as well. As situations arise where second-order derivatives are prohibitively expensive to compute, this version of CRNAS would be prudent to implement; therefore, for the sake of completeness, we include a description of our first-order version of CRNAS with associated convergence theory in Appendix A.3

Illustrative example of an ϵ-SOSP

Theoretical guarantees of convergence to an ϵ-SOSP are one of the main advantages of CRNAS. To facilitate understanding of the definition of an ϵ-SOSP, we present the following illustrative example. Consider the following optimization problem:

$$\begin{array}{ll}\min & f(\theta) := \sin(\frac{\pi}{2}\theta)\\ \text{s.t.} & -1 \leq \theta \leq 1.\end{array}$$

The derivatives are $f'(\theta) = \frac{\pi}{2}\cos(\frac{\pi}{2}\theta)$ and $f''(\theta) = -(\frac{\pi}{2})^2\sin(\frac{\pi}{2}\theta)$. At the same time, the log barrier function is $B(\theta) = -\ln(1-\theta) - \ln(1+\theta)$, so $B''(\theta) = \frac{2(1+\theta^2)}{(1-\theta^2)^2}$, which implies that the local norm is $\|d\|_\theta = \frac{\sqrt{2(1+\theta^2)}}{1-\theta^2}|d|$. In view of these, the ϵ-FOSP is

$$\left|\frac{\pi}{2}\cos(\frac{\pi}{2}\theta)\right| \leq \epsilon\frac{\sqrt{2(1+\theta^2)}}{1-\theta^2},$$

and the ϵ-SOSP is

$$-\frac{\pi^2}{4}\sin(\frac{\pi}{2}\theta) \geq -\sqrt{\epsilon}\frac{2(1+\theta^2)}{(1-\theta^2)^2}.$$

Noticing $\left|\frac{\pi}{2}\cos(\frac{\pi}{2}(1-\sqrt{\epsilon}))\right| = \left|\frac{\pi}{2}\sin(\frac{\pi}{2}\sqrt{\epsilon})\right| \approx O(\sqrt{\epsilon})$, it is evident that both $-1+\sqrt{\epsilon}$ and $1-\sqrt{\epsilon}$ qualify as $O(\epsilon)$-FOSP. However, only $-1+\sqrt{\epsilon}$ satisfies the $O(\epsilon)$-SOSP, which happens to be close to the global optimality in this case.

3. Results

3.1. Case study I: cancer drug response estimation

In the current and subsequent sections, we present two case studies on inferring heterogeneous population dynamics, rooted in real-world applications. In our case studies, we analyzed, evaluated, and compared the performance of CRNAS to state-of-the-art constrained nonlinear programming algorithms in MATLAB's optimization solver, fmincon [27]. This solver is widely used to solve

parameter estimation problems in mathematical biology [28–30]. Specifically, we compared CRNAS with the interior-point (IP and IP hess) and sequential quadratic programming (SQP and SQP grad) algorithms implemented in fmincon, where "hess" and "grad" indicate the use of analytic Hessian and gradient inputs, respectively. Details of candidate algorithms and comparison metrics are provided in Appendix A.4.

In this case study, we explore the use of CRNAS to solve a recently proposed parameter estimation problem [7, 31]. In these works, the authors employed a maximum likelihood estimation (MLE) approach to estimate tumor subpopulation dynamics and the corresponding drug response from high throughput drug screening data. The primary challenge in these studies arises from estimating the parameters of mixtures of dose-response Hill equations which are both nonlinear and non-convex. To alleviate some of the issues that arise from nonlinearity, we perform a parameter transformation on the variables in the Hill equations. Due to the importance and prevalence of the Hill equation, we review it before describing the cancer drug response estimation problem.

3.1.1. Hill equation and our parameter transformation

The Hill equation is a fundamental model to describe dose-response relationships in pharmacological studies [32]. For a given drug dose level d, the Hill equation specifies the fraction of viable cells as follows:

$$H(d; b, E, n) = b + \frac{1-b}{1+(\frac{d}{E})^n},$$

where the parameters b, E, and n represent the maximum drug effect, the half maximal effect dose (EC50), and the Hill coefficient, respectively. This model has been well utilized in pharmacology, biochemistry, and various other fields over the past hundred years [33].

An important question that arises in the study of the Hill equation is the accurate identification of model parameters based on observed data [34]. A particularly challenging aspect of dealing with the parameters of the Hill equation is the nonlinear term in the denominator, E^n. To mitigate the challenge of this term, we apply a variable transformation. In particular, we introduce a new parameter $\mathcal{E} = E^n$ to produce the following modified Hill equation:

$$\bar{H}(d; b, \mathcal{E}, n) = b + \frac{1-b}{1+\frac{d^n}{\mathcal{E}}}.$$

Given that the transformation $\mathcal{E} = E^n$ is a one-to-one function when E is non-negative, we can determine the value of E and n from the estimated values of $\mathcal{E}$ and n, respectively. This parameter transformation is applied during the numerical experiments, while the original EC50 parameter E is used throughout the manuscript for clarity.

3.1.2. Deterministic drug-affected cell proliferation framework

In [31], the authors proposed a novel statistical framework, PhenoPop, to analyze high-throughput drug screening data. PhenoPop takes the tumor drug response into account when multiple subpopulations, each with varying drug responses, exist within the tumor. A single Hill equation cannot capture this heterogeneous drug response; therefore, the PhenoPop model represents the dynamics of tumor growth as a composite of numerous subpopulations, with each subpopulation distinguished by a unique set of Hill parameters.

PhenoPop employs a classical exponential growth model to describe the growth of each subpopulation. Each subpopulation has a distinct growth rate, denoted as α_i. The size of the subpopulation i at time t is represented as follows:

$$X_i(t) = X_i(0)\exp(\alpha_i t),$$

where $X_i(0)$ is the initial population size of subpopulation i. The unique set of Hill parameters for subpopulation i is denoted as (b_i, E_i, n_i). The relationship between the cell growth rate and the drug dose is expressed as follows:

$$\alpha_i(d; b_i, E_i, n_i) = \alpha_i + \log(H(d; b_i, E_i, n_i)),$$

where the constraint $b_i \in (0, 1)$ is enforced to ensure that the drug decreases the growth rate. For $i \in \{1, \cdots, S\}$, the cell count of subpopulation i at time t and dose levels d is modeled as follows:

$$f_i(t, d) = p_i X(0)\exp[t(\alpha_i + \log(H(d; b_i, E_i, n_i)))], \tag{3.1}$$

where $p_i \in (0, 1)$ satisfies $\sum_i p_i = 1$ and represents the initial proportion of subpopulation i, and $X(0)$ is the initial total cell count. While $X(0)$ is assumed to be a known quantity, PhenoPop estimates the following set of parameters from the data:

$$\theta_{PP}(S) = \{p_i, \alpha_i, b_i, E_i, n_i; i \in \{1, \cdots, S\}\}.$$

To estimate these parameters, cell counts are collected at a set of time points $t \in \{t_1, \cdots, t_T\} = \mathcal{T}$ and drug dose levels $d \in \{d_1, \cdots, d_D\} = \mathcal{D}$; R replications are performed for each possible time-dose combination. We denote the dataset as follows:

$$\mathcal{X} = \{x_{t,d,r}; t \in \mathcal{T}, d \in \mathcal{D}, r \in \{1, \cdots, R\}\}.$$

Then, PhenoPop estimates the parameter set through the MLE process. In [31], the authors assumed Gaussian noise with a mean of zero, featuring two possible variances based on the dose level and time of observation. For simplicity, we assume that the noise follows a mean-zero Gaussian distribution with a constant variance σ^2. The corresponding negative log-likelihood function for the observations $\mathcal{X}$ under the parameters set θ is given by the following:

$$-\log(L(\theta, \sigma^2; \mathcal{X})) = -DRT\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{1}{2\sigma^2}\sum_{t\in\mathcal{T}}\sum_{d\in\mathcal{D}}\sum_{r=1}^{R}(x_{t,d,r} - f(t, d; \theta))^2, \tag{3.2}$$

where $f(t, d; \theta) = \sum_{i=1}^{S} f_i(t, d)$. In the *in silico* validation of CRNAS, we assume that the observation noise σ^2 is known. Therefore, the MLE problem is reduced to the following constrained least square problem:

$$\hat{\theta} \in \arg\min_{\theta\in\Theta}\sum_{t\in\mathcal{T}}\sum_{d\in\mathcal{D}}\sum_{r=1}^{R}(x_{t,d,r} - f(t, d; \theta))^2, \tag{3.3}$$

where $\hat{\theta}$ is an estimated parameter set, and Θ is the feasible space of the parameters. In practice, the feasible space for the parameters is selected based on several biological assumptions. We provide detailed specifications of the feasible space for each experiment in Appendix A.5.

Numerical results:

To evaluate the performance of each algorithm to solve (3.3), we conducted a series of *in silico* experiments. In each experiment, we generated 100 different datasets, each corresponding to a randomly selected true parameter set. For each dataset, we solved (3.3) starting from 20 different initial points. Among those 20 results, we recorded the result with the best objective value as the solution obtained from each algorithm. Details regarding the problem initialization and data simulation for these tests are located in Appendix A.5. It is important to note that in PhenoPop, the primary source of randomness is the observation noise, which is less significant in this context; therefore, for the simulated data, we assume there is no observation noise, thereby aiming to better illustrate each algorithms' performance. Thus, the optimal objective value should be close to zero; we consider values that exceed one indicative of poor estimation. Our discussion of the experiments is divided into three sections based on the number of assumed subpopulations, S, in the model.

Experiment with $S = 1$ (*in silico*)

We begin with a preliminary experiment that focuses on tumor dynamics that involve a single dose-response curve. In this case, the parameter set consists of α, b, E, and n, without specification of the initial proportion parameter p. Figure 1 displays the three performance metrics discussed in Appendix A.4 for each of the five algorithms listed in Table A1. The box-plots display the results from 100 unique and independent datasets generated for the PhenoPop model.

By evaluating each algorithm with our performance metrics, we see that CRNAS demonstrated a superior performance compared to the other algorithms in terms of the number of iterations required to achieve its best solution from one initial guess. When comparing the total compute times, it is evident that CRNAS required less time compared to IP and IP hess, while SQP and SQP grad required a similar amount of time to run for 20 different initial guesses; however, though SQP and SQP grad had comparable total compute times, they yielded less reliable solutions. CRNAS consistently delivered accurate estimates across all but one of the 100 instances of the model, as evidenced by its optimal obtained objective values consistently falling significantly below one in all but one test. On the other hand, both SQP and SQP grad produced multiple poor estimations. Notably, without higher-order information, both SQP and IP could not reduce the objective value below 10^{-8}. This demonstrates a handicap of utilizing the inexact higher-order information. With the exact higher-order information, both SQP grad and IP hess yielded optimal solutions lower than their inexact higher-order counterparts. One might argue that since SQP grad has a comparable estimation quality to CRNAS, statistically speaking, as evidenced by Figure 1, there is limited advantage to using CRNAS; however, this would be fallacious given the fact that SQP grad produced poor estimates in about ten percent of the tests. IP hess produced accurate estimates, but required more time to obtain these solutions compared to CRNAS. Overall, we see CRNAS performs on par, if not better than, MATLAB's solvers based on our metrics.

Experiment with $S = 2$ (*in silico*)

Now, we consider the case where two subpopulations exist in the tumor, each with a distinct drug response. Therefore, the initial proportion parameter p_i for each subpopulation is reintroduced along with the corresponding equality constraint, $p_1 + p_2 = 1$; Figure 2 provides the results for these experiments.

With two subpopulations present, CRNAS maintained its advantage in terms of the number of iterations needed to obtain the optimal solution compared to the other algorithms. Adding an additional

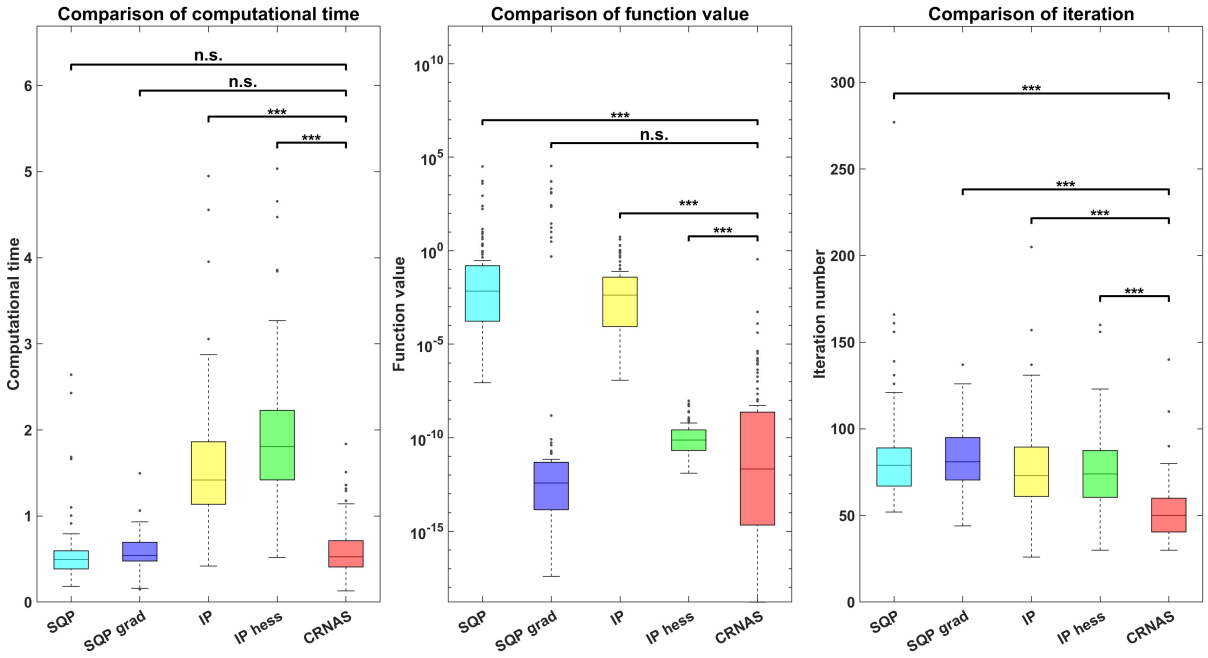

Figure 1. Comparison of the algorithms in Table A1 to solve (3.3) with a single subpopulation (i.e., $S = 1$). The left panel records each algorithm's total compute time to solve each of the 100 instantiations of the model from 20 different initial points. The middle panel records the best function value obtained by each algorithm from the 20 differential initial points for each of the 100 different experiments; the Y axis is in a logarithmic scale. The right panel records the number of iterations taken to obtain the optimal solution that corresponds to the best objective values displayed in the middle panel. Each boxplot indicates the upper extremes, upper quantile, median, lower quantile, and lower extremes; the scattered dots represent outliers. The significance bars indicate the p-values derived from the Wilcoxon rank sum test with significance levels such that $*** \leq 0.001 \leq ** \leq 0.01 \leq * \leq 0.05 \leq n.s.$.

subpopulation increased the overall complexity of the model and this added complexity greatly limited SQP and SQP grad's ability to obtain accurate solutions. In more than 1/4 of the tests, SQP's optimal computed objective values exceeded 1, while more than 3/4 of the optimal solutions obtained by SQP grad were worse than the median of the optimal objective values obtained by CRNAS. In contrast, CRNAS provided accurate estimations across all experiments for this more challenging problem; the performances of IP and IP hess were similar to the prior experiment.

To provide more comprehensive numerical results, we studied the quality of the results when all proposed algorithms were allocated the same amount of time. Specifically, we restricted the computational time for each algorithm to be 1 second. Within 1 second, each algorithm can try as many initial guesses as possible.

Then, we conducted 100 random experiments, each generated by one randomly selected true parameter set. To assess the quality of solutions, we define any solution with an objective value below threshold of 1 as a good solution. The number of objectives below a specific threshold value of 1, and the average number of initial trials among the 100 experiments conducted by each algorithm, are

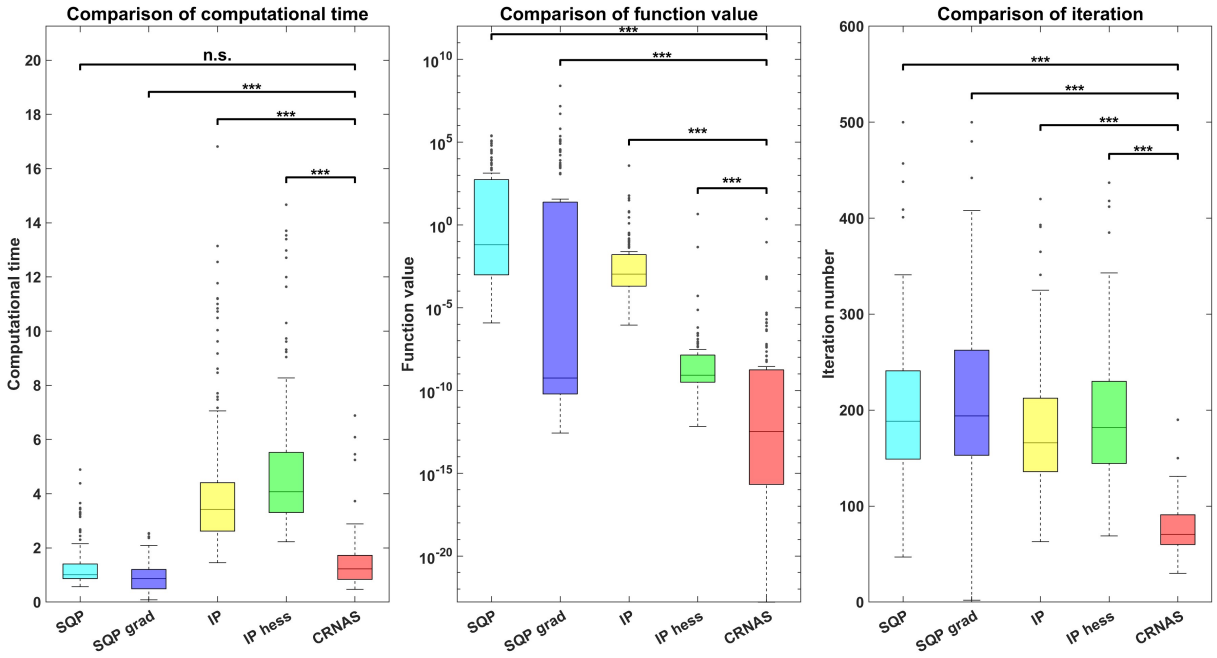

Figure 2. Comparison of the algorithms in Table A1 to solve (3.3) with $S = 2$ subpopulations; no upper bound constraint was present for E_i and n_i in these experiments. The results are presented as before in Figure 1.

presented in Table 1. Within the same time restriction, CRNAS found the objective below the threshold in more experiments than other algorithms. In terms of the number of initial trials, CRNAS roughly tried the same number of initial points as SQP and fewer than SQP grad, while all three algorithms tried more than IP and IP hess. These results are consistent with the observations from our previous experiments with a fixed number of initial trials, where the CRNAS demonstrated a similar time-efficiency performance to SQP, a worse performance than SQP grad, and a better performance than IP and IP hess.

Table 1. Comparison of the algorithms in Table A1 to solve (3.3) with $S = 2$ subpopulations when computational time is fixed.

Algorithms	Number of experiments that have objective below 1	Average initial trials
SQP	58	24
SQP grad	82	44
IP	86	9
IP hess	94	7
CRNAS	95	24

In seeking to understand the factors which limited the performance of SQP, we found that imposing an upper bound constraint on some of the parameters improved the performance of SQP. This technique requires researchers to heuristically determine an upper bound for all parameters. If the true parameter set is located within the narrowed zone, then this significantly restricts the search

region for the optimization algorithm and aids its location of better estimates. The risk of employing this practice depends on the accuracy of the practitioner's prior knowledge, since placing too small of an upper bound could make finding the true global optimal impossible as it will be outside the artificially constructed feasible region. For the in silico experiments, we have full knowledge about the true parameter set $\Theta^*(S)$, thus leading to an accurate optimization feasible region selection. Specifically, we set the optimal feasible region for E_i and set n_i to be the interval $(0, 100)$. The constraints on the other parameters remained the same.

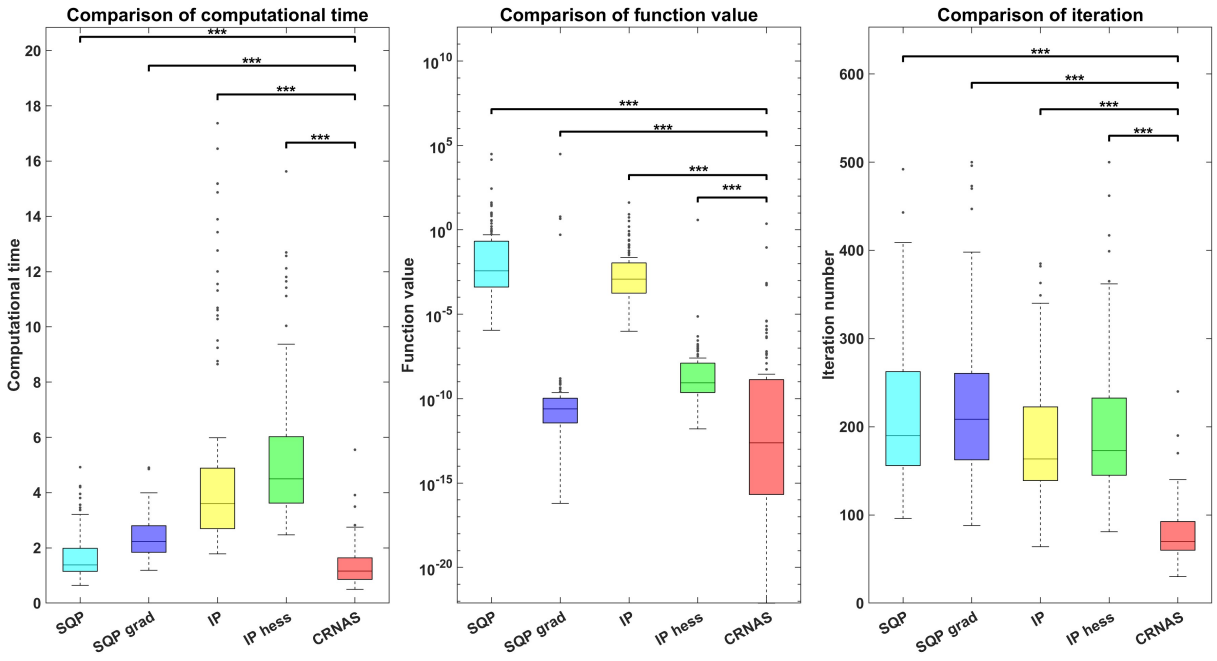

Figure 3. Comparison of the algorithms in Table A1 to solve (3.3) with $S = 2$ subpopulations. The "box constraint" is applied to each parameter. The results are presented as in Figure 1.

Another set of experiments was conducted using the new upper bound constraints on E_i and n_i; the results are provided in Figure 3. With these constraints, SQP grad performed in a similar fashion to the experiments with one subpopulation; however, it required a greater total compute time compared to the outcomes depicted in Figure 2. Consequently, CRNAS significantly outperformed the other methods in both the total compute time and the number of iterations. Additionally, CRNAS can provide robust estimates even without specifying upper bounds for all parameters, as shown in Figure 2. Therefore, a practitioner does not need to estimate potentially unknown upper bounds on parameters for their models in order for CRNAS to obtain strong parameter estimates.

Experiment with $S > 2$ (*in silico*)

Lastly, we examined two more cases with $S = 3$ and $S = 5$. The majority of the conditions to generate the data remain unchanged from the experiments with $S = 2$; however, to accommodate setting $S = 3$ and $S = 5$, we employed a new scheme to select the EC50 values for each subpopulation and the dose levels $\mathcal{D}$ to ensure the identifiability of each parameter set; for details, see Appendix A.6.

The results for $S = 3$ and $S = 5$ are shown in Figures 4 and 5, respectively. When $S = 3$, it is

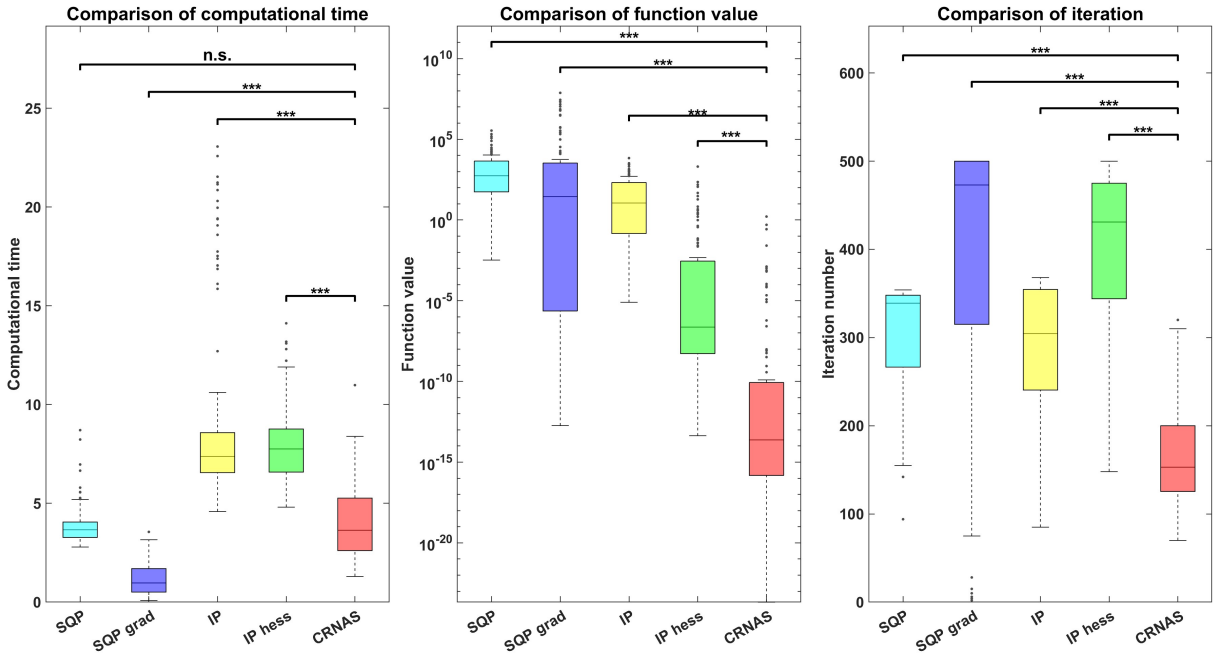

Figure 4. Comparison of the algorithms in Table A1 to solve (3.3) with $S = 3$ subpopulations. The results are presented as in Figure 1.

evident that IP hess, which produced very accurate solutions in the prior experiments, located almost zero accurate estimates with nearly all its computed optimal objective values being larger than one. Some of the poor estimates produced by IP hess and SQP grad might be due to the algorithms that reach the maximum iteration limit of 500 (see Appendix A.8 for details on termination criteria); however, the algorithms produced poor estimates even when the iteration limit was not reached, as supported by the fact that SQP grad and IP hess terminated before the maximum number of iterations was reached in more than half of the experiments with three subpopulations. However, CRNAS consistently provided accurate estimates for both three and five subpopulations. We propose that this advantage stems from the inherent capability of CRNAS to automatically avoid saddle points and converge to second-order stationary points, thereby avoiding suboptimal first-order stationary points which the other methods are potentially converging to.

Experiment using real-world data (*in vitro*)

To showcase the applicability of CRNAS to real-world applications, we applied it to fit an *in vitro* dataset generated by [31]. This dataset consists of four experimental datasets (BF12, BF11, BF21, and BF41), in which imatinib-sensitive and imatinib-resistant *Ba/F3* cells were mixed and cultivated at predefined ratios. To capture the dynamics of the sensitive and resistant populations, we employed a deterministic model that assumed $S = 2$. Unlike the datasets used in the *in silico* experiments, *in vitro* datasets are subject to noise from multiple sources, including counting errors and intrinsic stochasticity in cellular dynamics. Therefore, we reintroduced the observation noise parameter σ and jointly minimized the negative log-likelihood in Eq (3.2) with respect to θ and σ. Due to the presence of observation noise, the optimal parameter values at the global optimum are no longer equal to zero. Accordingly, we define the global optimums as those obtained from multiple optimization trials.

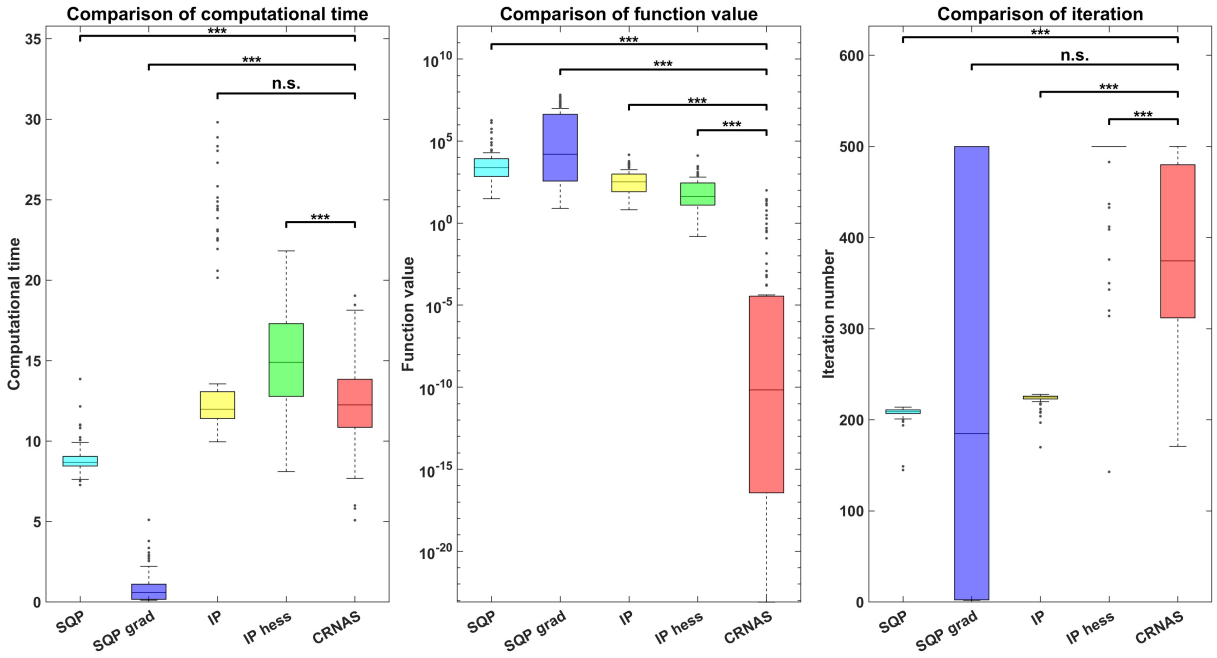

Figure 5. Comparison of the algorithms in Table A1 to solve (3.3) with $S = 5$ subpopulations. The results are presented as in Figure 1.

Table 2 reports the global optimums and the number of iterations required to fit the *in vitro* datasets using the corresponding optimization algorithms. Across twenty trials, all five algorithms converge to comparable optima. Importantly, CRNAS converges to the global optimum in fewer iterations, which is consistent with our previous observations from the *in silico* experiments. This behavior indicates that CRNAS achieves the global optimum with a lower computational cost.

To assess the robustness of CRNAS in fitting real-world data, we examined all the solutions obtained from these twenty trials. The ordered objective values in Figure 6 indicates that CRNAS exhibits the best robustness to initial conditions, with more than fifteen out of twenty trials converging to solutions close to the global optimum. In contrast, more than eight of the twenty trials from the other algorithms converge to objective values that are 10^2 higher than the global optimum. In addition to its robust performance, CRNAS requires a total computational time comparable to that of the other algorithms. In particular, CRNAS required a smaller total run time to complete all twenty trials than IP and IP Hess. Although both SQP and SQP Grad required smaller runtimes than CRNAS, they exhibited poor robustness to initial conditions. These two algorithms stop early when the objective values are not sufficiently close to the global optimum. According to the stopping messages, these algorithms terminated because the first-order optimality condition measure fell below the specified tolerance in multiple trials. This observation highlights the importance of seeking points that satisfy not only the first-order but also the second-order optimality conditions when the objective function is non-convex.

3.1.3. Stochastic drug-affected cell proliferation framework

To more accurately represent the changing variability in the data, [7] extended the PhenoPop framework using linear birth-death processes to model the heterogeneous tumor cell populations.

Table 2. Global optimums and corresponding iteration required for each algorithm in fitting the *in vitro* data (BF12|BF11|BF21|BF41).

Algorithms	Global optimums (twenty trials)	Iteration
CRNAS	(15490|14185|14346|16134)	(105|105|111|122)
SQP	(15490|14185|14346|16134)	(152|232|148|205)
SQP grad	(15490|14185|14346|16134)	(206|260|227|194)
IP	(15490|14185|14346|16134)	(261|275|169|311)
IP hess	(15490|14185|14346|16134)	(223|280|204|282)

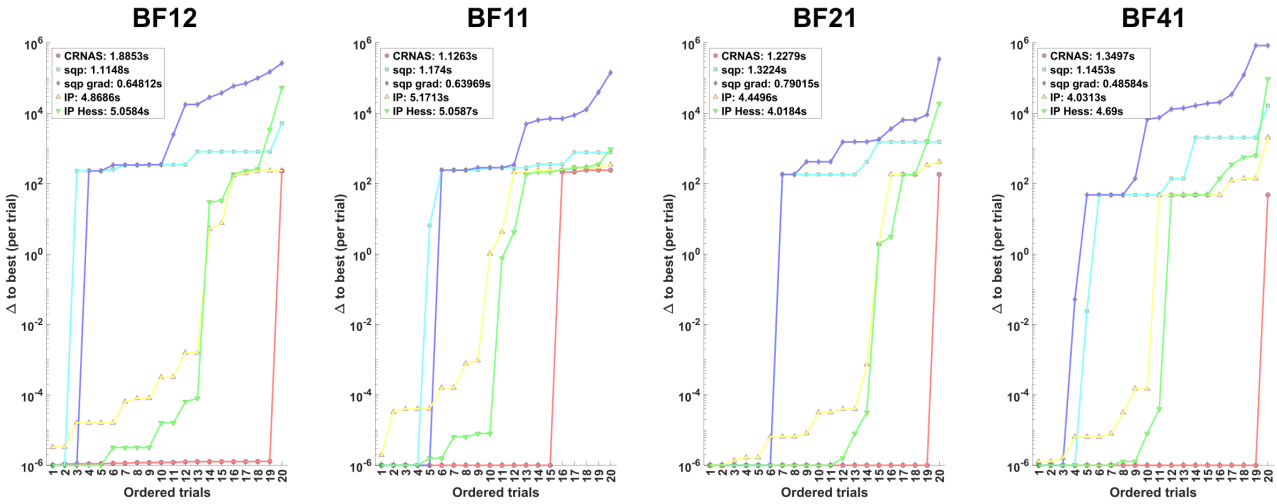

Figure 6. Line plots of the algorithms performance from twenty random trials in fitting the *in vitro* dataset. The ordered objective values are plotted as the difference to the global optimum, with a threshold 10^{-6}. Total computational time required for each algorithm are shown in the legend.

Instead of using a deterministic exponential growth model to represent the cell proliferation dynamics, they utilized the stochastic linear birth-death process. As a result, the relationship between the variance and the number of cells in the linear birth-death process can easily account for changing noise levels in the observations. Specifically, the authors assumed a cell in subpopulation i (type-i cell) stochastically divides into two cells at a rate of $\beta_i \geq 0$ and dies at a rate of $\nu_i \geq 0$. This means that during a short time interval $\Delta t > 0$, a type-i cell divides with a probability of $\beta_i \Delta t$ and dies with a probability of $\nu_i \Delta t$. Then, the drug effect on type-i cells is modeled as a cytotoxic effect as follows:

$$\nu_i(d) = \nu_i - \log(H(d; b_i, E_i, n_i)).$$

Note that this framework can also easily account for a cytostatic effect. The authors denoted the stochastic process $X_i(t, d)$ as the number of cells in the subpopulation i at time t under drug dose d with mean and variance

$$\mathbb{E}[X_i(t, d)] := X(0) p_i \mu_i(t, d) = X(0) p_i \exp(t(\beta_i - \nu_i(d)))$$

and

$$\mathrm{Var}[X_i(t,d)] := X(0)p_i\sigma_i^2(t,d) = X(0)p_i\frac{\beta_i+\nu_i(d)}{\beta_i-\nu_i(d)}\left(\exp(2t(\beta_i-\nu_i(d))) - \exp(t(\beta_i-\nu_i(d)))\right),$$

respectively. Since each subpopulation was assumed to independently grow, the total cell count $X(t,d) = \sum_{i=1}^{S} X_i(t,d)$ is normally distributed with mean and variance

$$\mathbb{E}[X(t,d)] := \mu(t,d) = \sum_{i=1}^{S} X(0)p_i\mu_i(t,d)$$

$$\mathrm{Var}[X(t,d)] := \sigma^2(t,d) = \sum_{i=1}^{S} X(0)p_i\sigma_i^2(t,d),$$

respectively, when the initial population $X(0)$ is large. Consequentially, the authors suggested the following statistical framework:

$$x_{t,d,r} = X^{(r)}(t,d) + Z_{t,d,r},$$

where $X^{(r)}(t,d)$ are independent and identical distributed (i.i.d.) copies of $X(t,d)$ for $r = 1,\cdots,R$, and $\{Z_{t,d,r}; d \in \mathcal{D}, t \in \mathcal{T}, r \in \{1,\cdots,R\}\}$ are i.i.d. normally distributed observation noise with a mean of 0 and a variance of c^2. Thus, the model parameter set is now as follows:

$$\theta_{LBD}(S) = \{p_i, \beta_i, \nu_i, b_i, E_i, n_i, c; i \in \{1,\cdots,S\}\}.$$

This novel statistical framework aims to utilize the information in the variability of the data to predict the parameters $p_i, \beta_i, \nu_i, b_i, E_i$, and n_i related to each subpopulation. As a result, there is a more complicated negative log-likelihood for the MLE process:

$$-\log(L(\theta_{LBD}(S);\mathcal{X})) = -\sum_{t,d\in\mathcal{T},\mathcal{D}} R\log\left(\frac{1}{\sqrt{2\pi(\sigma^2(t,d)+c^2)}}\right) + \sum_{t,d\in\mathcal{T},\mathcal{D}}\sum_{r=1}^{R}\frac{(x_{t,d,r}-\mu(t,d))^2}{2(\sigma^2(t,d)+c^2)}. \tag{3.4}$$

The challenge in minimizing the negative log-likelihood function in (3.4) is its variance term. Now, the variance depends on the cell growth dynamics of each subpopulation which involves the Hill equation.

Numerical results:

Next, we investigate CRNAS's performance on the MLE problem described by (3.4). Details for the MLE problem initialization, optimization feasible range Θ, and data simulation are provided in Appendix A.6. In the stochastic model, verifying whether the solution reaches the global optimum is challenging, particularly when the parameter space is infinite. To address this, we use the likelihood of the true parameter set $\theta^*_{LBD}(S)$ as a targeted likelihood value. Specifically, we consider a solution to be good if its likelihood exceeds that of the true parameter set. In other words, a good solution should have a lower objective value (negative log-likelihood) than that of the true parameter set. Since the likelihood of the true parameter $\theta^*_{LBD}(S)$ varies across different experiments, in order to create a unified metric across experiments, we employed the relative likelihood for an estimator $\hat{\theta}_{LBD}(S)$ defined as follows:

$$RL(\hat{\theta}_{LBD}(S);\theta^*_{LBD}(S)) := \frac{-\log(L(\hat{\theta}_{LBD}(S);\mathcal{X}(\theta^*_{LBD}(S))))}{-\log(L(\theta^*_{LBD}(S);\mathcal{X}(\theta^*_{LBD}(S))))}, \tag{3.5}$$

where the likelihood function $L(\cdot; \mathcal{X})$ is defined as in (3.4), and $\mathcal{X}(\theta^*_{LBD}(S))$ is the dataset generated by the true parameter set $\theta^*_{LBD}(S)$; typically this ratio is below 1 for quality estimators.

As discussed in Section 3.1.3, the advantage and novelty of the linear birth-death process framework is that its variance structure can naturally explain the dynamic variance observed in real-world data. With mild assumptions on the cell growth dynamics, the linear birth-death process framework is able to utilize the information in the variance of the data to obtain more robust estimates of the model's parameters. However, the trade-off is that employing this variance structure further complicates the likelihood function, as the variance term also contains a Hill function. For simplicity, we consider the setting with $S = 2$ subpopulations. In Section 3.1.2, we concluded that inexact gradient and Hessian computations may mislead IP and SQP to terminate at poor solutions. Thus, we focus on comparing IP hess, SQP grad, and CRNAS in these experiments; Figure 7 displays the results of the experiments.

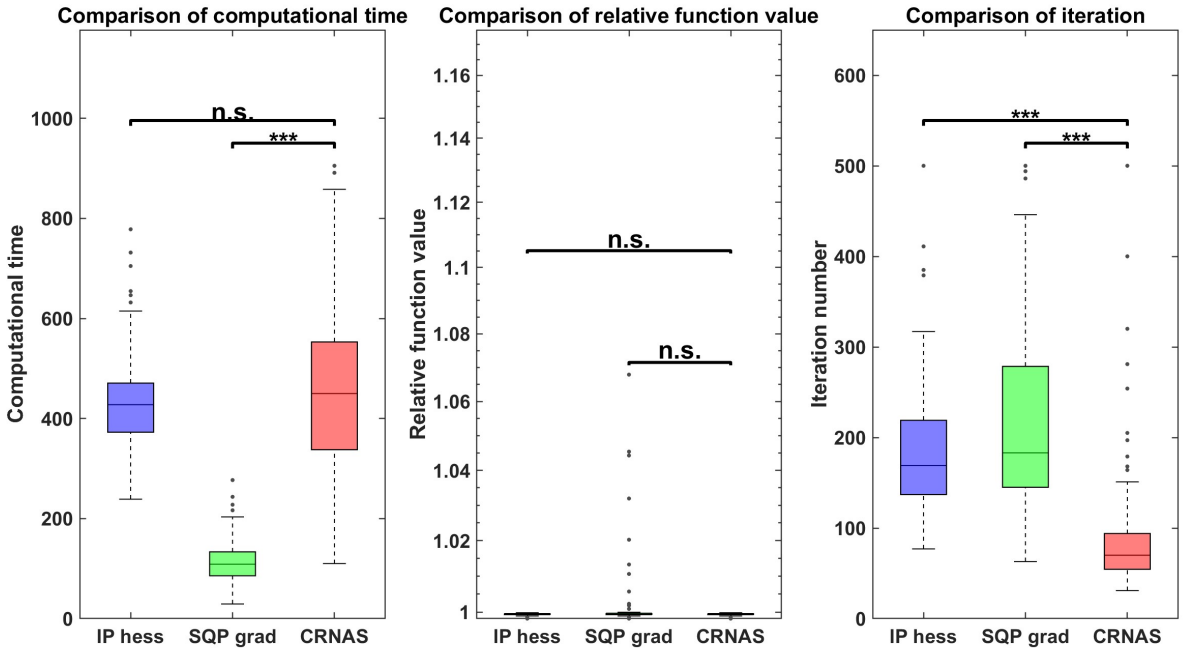

Figure 7. Comparison of three optimization algorithms to solve the MLE problem in (3.4) with $S = 2$ subpopulations. The results are presented as in Figure 1.

From the results, we note that CRNAS still has the best convergence rate in terms of the number of iterations; however, the total compute time of CRNAS was comparable to IP hess and longer than SQP grad. This is primarily because, for this problem, evaluating the Hessian matrix is significantly more expensive than evaluating the gradient. Though SQP grad solved the problem relatively quickly, many of the solutions were of low quality. Specifically, 14 of the 100 solutions obtained by SQP grad had a relative likelihood greater than 1, while all of the estimations obtained from IP hess and CRNAS had a relative likelihood less than 1. To illustrate this, we present an example estimation in Table 3. In this example, it is evident that SQP grad fails to accurately estimate the parameters for the second subpopulation, as the estimations for E_2 and n_2 are completely incorrect. This example reiterates the necessity of including the exact computation of higher-order information for better-quality solutions.

Additionally, it supports our claim that when the relative likelihood is below 1, the estimation quality is high.

Table 3. An illustrative example of the results in Figure 7. The dataset was generated based on the true parameter set $\theta^*(2)$, and the estimations of IP hess, SQP grad, and CRNAS are shown. SQP grad incorrectly estimated both E_2 and n_2; these estimates are bolded. *RL* in the last column is the relative likelihood defined in (3.5).

	p_1	β_1	ν_1	b_1	E_1	n_1	p_2	β_2	ν_2	b_2	E_2	n_2	*RL*
$\theta^*(2)$	0.36	0.41	0.31	0.8	0.08	2.05	0.64	0.72	0.67	0.95	0.81	4.39	1
IP hess $\hat{\theta}(2)$	0.35	0.42	0.33	0.79	0.09	1.96	0.65	0.71	0.67	0.95	0.83	4.42	0.9997
SQP grad $\hat{\theta}(2)$	0.38	0.43	0.34	0.8	0.09	1.82	0.62	0.73	0.68	0.95	**1.19**	**36.51**	1.0009
CRNAS $\hat{\theta}(2)$	0.35	0.42	0.33	0.79	0.09	1.96	0.65	0.71	0.67	0.95	0.83	4.42	0.9997

3.2. Case study II: heterogeneous logistic estimation

The logistic growth model is popular to study population growth with a carrying capacity. Recently, there has been significant interest in an extended version of the logistic growth model, which aims to represent cell populations that consist of heterogeneous subpopulations [35]. These subpopulations can exhibit varying growth dynamics, such as differing growth rates and carrying capacities. Here, for simplicity, we study a linear combination of logistic growth functions with varying parameter sets for each subpopulation to model a mixture of heterogeneous populations growing according to a logistic function. Note that we do not assume the populations interact via their carrying capacity, as this would necessitate the numerical solution of a nonlinear system of differential equations, which is outside the focus of this manuscript.

In particular, we start with the parameters of S subpopulations: $\theta_{LG}(S) = (p_i, \alpha_i, \beta_i), i = 1, \cdots, S$. Then, we model the total cell count dynamics with respect to time as follows:

$$F_{LG}(t; \theta_{LG}(S)) = \sum_{i=1}^{S} f_i(t; p_i, \alpha_i, \beta_i) = \sum_{i=1}^{S} F_{LG}(0) p_i \frac{1}{1 + e^{-\alpha_i t + \beta_i}}, \tag{3.6}$$

where $F_{LG}(0)$ is the initial total cell count, and p_i is the initial proportion of each subpopulation, thereby satisfying an equality constraint $\sum_{i=1}^{S} p_i = 1$. Compared to the widely used three-parameter logistic growth model, we normalize the maximum carrying capacity for each subpopulation to 1, thus ensuring the identifiability of each parameter. We retain another two parameters, (α_i, β_i), to capture the growth behavior of a type i population at the inflection point. The parameter α_i encodes the maximum growth rate that the type i population can achieve around the inflection point, while the parameter β_i relates to the shift in time of the inflection point. From a modeling perspective, there are no restrictions on α_i and β_i. However, we assume $\alpha_i \geq 0$ for all $i = 1, \cdots, S$ for a growing population and select a distinct β_i for each subpopulation to ensure that these parameters are identifiable. The detailed selection method is specified in Appendix A.7. For this experiment, we assume a zero observation noise. Thus, for a dataset observed at a given set of time points $\mathcal{T}$, $\mathcal{X} = \{x_t; t \in \mathcal{T}\}$, we solve the least squares problem

$$\hat{\theta}(S) = \underset{\theta(S) \in \Theta(S)}{\arg\min} \sum_{t \in \mathcal{T}} (x_t - F_{LG}(t; \theta(S)))^2 \tag{3.7}$$

to estimate the true parameter set $\theta^*_{LG}(S)$. The detailed feasible space $\Theta(S)$ for the experiment is provided in Appendix A.7.

Numerical results:

We compared the performance of all the algorithms in Table A1 to solve the least squares problem described in (3.7). Similar to the previous case study, we compared the methods based on the described metrics in Appendix A.4. Experimental details are provided in Appendix A.7.

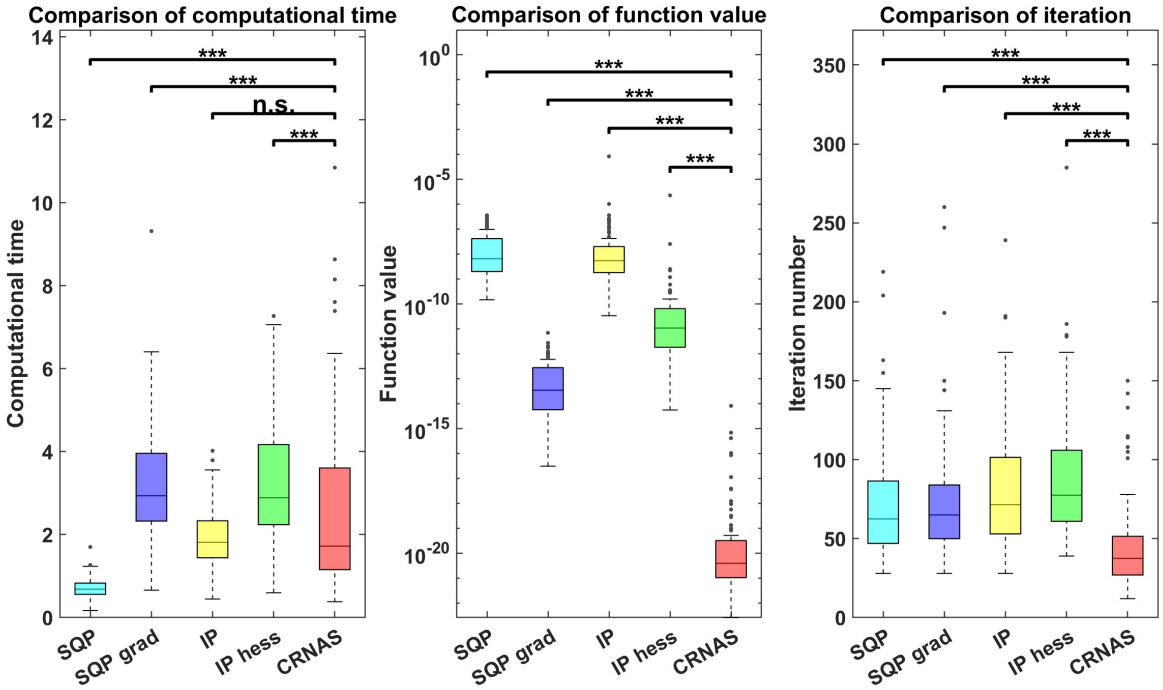

Figure 8. Comparison of the algorithms in Table A1 to solve the least square problem in (3.7) with $S = 2$ subpopulations. The results are presented as in Figure 1.

From Figure 8, we see again that CRNAS boasted the smallest number of iterations required to obtain the optimal solution. Furthermore, the middle panel in Figure 8 shows that CRNAS yielded the lowest best objective values among all five methods. Based on the advantages observed in both the number iterations and the objective value, we see that CRNAS outperformed four different implementations of state-of-the-art algorithms to solve this highly nonlinear and non-convex parameter estimation problem.

4. Conclusions

This paper introduced a novel optimization algorithm called CRNAS to solve nonlinear and non-convex constrained optimization problems. CRNAS combines the best qualities of Nesterov's cubic regularized Newton's method and the classical affine scaling method for linear programming to produce a new approach with convergence guarantees to second-order stationary points. The design of our algorithm was primarily motivated by the parameter estimation problems that arise in modeling heterogeneous population dynamics, where parameters must strictly adhere to predefined box and linear equality constraints. These problems produce exceptionally challenging constrained

optimization models, and our results demonstrated that CRNAS competes with state-of-the-art algorithms.

We conducted a comparative analysis between CRNAS and two state-of-the-art local optimization methods implemented in MATLAB's commercial solver fmincon. This analysis was performed across two case studies, and our findings consistently demonstrated that CRNAS outperformed *fmincon* by generally delivering more accurate solutions, requiring fewer iterations, and achieving comparable, if not better, computational times; however, the computational cost of evaluating second-order information may diminish the advantage of CRNAS. Since CRNAS requires computing the Hessian at each iteration, we recommend its use when second-order information can be efficiently obtained and at a comparable cost to first-order derivatives.

Nevertheless, in mathematical biology, Hessian information is often accessible with high accuracy, as the objective is typically available in an analytical form, and the problem dimension are generally moderate. A key direction for future work is extending CRNAS to a broader class of parameter estimation problems in mathematical biology. One particularly promising application is parameter estimation in ordinary differential equation (ODE) models, where both gradients and Hessians can be computed via sensitivity analysis with controllable precision.

Authorship contribution statement

Chenyu Wu: Conceptualization, Investigation, Software, Validation, Visualization, Writing-original draft, Writing-review & editing. **Nuozhou Wang:** Investigation, Methodology, Software, Validation, Writing-original draft, Writing-review & editing. **Casey Garner:** Methodology, Writing-review & editing. **Kevin Leder:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing-review & editing. **Shuzhong Zhang:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing-review & editing.

Code availability

The code necessary to reproduce the experimental results presented in this paper are published in https://github.com/chenyuwu233/CRNAS-Experiment-code.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgements

The work of K. Leder, S. Zhang, C. Wu and N. Wang was partially supported by NSF Award CMMI 2228034; C. Garner is supported by the NSF-GRFP under Grant No. 2237827.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. H. Hass, C. Loos, E. Raimúndez-Álvarez, J. Timmer, J. Hasenauer, C. Kreutz, Benchmark problems for dynamic modeling of intracellular processes, *Bioinformatics*, **35** (2019), 3073–3082. https://doi.org/10.1093/bioinformatics/btz020
2. Y. Cao, S. Li, L. Petzold, R. Serban, Adjoint sensitivity analysis for differential-algebraic equations: The adjoint DAE system and its numerical solution, *SIAM J. Sci. Comput.*, **24** (2003), 1076–1089. https://doi.org/10.1137/S1064827501380630
3. E. G. Birgin, J. M. Martínez, On regularization and active-set methods with complexity for constrained optimization, *SIAM J. Optim.*, **28** (2018), 1367–1395. https://doi.org/10.1137/17M1127107
4. E. G. Birgin, G. Haeser, A. Ramos, Augmented Lagrangians with constrained subproblems and convergence to second-order stationary points, *Comput. Optim. Appl.*, **69** (2018), 51–75. https://doi.org/10.1007/s10589-017-9937-2
5. A. Daigeler, L. Klein-Hitpass, A. M. Chromik, O. Müller, J. Hauser, H. H. Homann, et al., Heterogeneous in vitro effects of doxorubicin on gene expression in primary human liposarcoma cultures, *BMC Cancer*, **8** (2008), 1–17. https://doi.org/10.1186/1471-2407-8-313
6. M. Nouiehed, J. D. Lee, M. Razaviyayn, Convergence to second-order stationarity for constrained non-convex optimization, preprint, arXiv:1810.02024. https://doi.org/10.48550/arXiv.1810.02024
7. C. Wu, E. B. Gunnarsson, E. M. Myklebust, A. Köhn-Luque, D. S. Tadele, J. M. Enserink, et al., Using birth-death processes to infer tumor subpopulation structure from live-cell imaging drug screening data, *PLoS Comput. Biol.*, **20** (2024), e1011888. https://doi.org/10.1371/journal.pcbi.1011888
8. L. Ljung, T. Chen, Convexity issues in system identification, *2013 10th IEEE Int. Conf. Control Autom. (ICCA)*, (2013), 1–9. https://doi.org/10.1109/ICCA.2013.6565206
9. C. G. Moles, P. Mendes, J. R. Banga, Parameter estimation in biochemical pathways: A comparison of global optimization methods, *Genome Res.*, **13** (2003), 2467–2474. https://doi.org/10.1101/gr.1262503
10. L. Schmiester, D. Weindl, J. Hasenauer, Efficient gradient-based parameter estimation for dynamic models using qualitative data, *Bioinformatics*, **37** (2021), 4493–4500. https://doi.org/10.1093/bioinformatics/btab512
11. A. Gábor, J. R. Banga, Robust and efficient parameter estimation in dynamic models of biological systems, *BMC Syst. Biol.*, **9** (2015), 1–25. https://doi.org/10.1186/s12918-015-0219-2
12. F. Fröhlich, P. K. Sorger, Fides: Reliable trust-region optimization for parameter estimation of ordinary differential equation models, *PLoS Comput. Biol.*, **18** (2022), e1010322. https://doi.org/10.1371/journal.pcbi.1010322
13. P. Stapor, F. Fröhlich, J. Hasenauer, Optimization and profile calculation of ode models using second order adjoint sensitivity analysis, *Bioinformatics*, **34** (2018), i151–i159. https://doi.org/10.1093/bioinformatics/bty230

14. N. Boumal, V. Voroninski, A. Bandeira, The non-convex Burer-Monteiro approach works on smooth semidefinite programs, in *30th Conference on Neural Information Processing Systems (NIPS 2016)*, (2016).

15. R. Ge, J. D. Lee, T. Ma, Matrix completion has no spurious local minimum, in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, (2016).

16. F. E. Curtis, D. P. Robinson, M. Samadi, An inexact regularized Newton framework with a worst-case iteration complexity of for nonconvex optimization, *IMA J. Numer. Anal.*, **39** (2019), 1296–1327. https://doi.org/10.1093/imanum/dry022

17. M. O'Neill, S. J. Wright, A log-barrier Newton-CG method for bound constrained optimization with complexity guarantees, *IMA J. Numer. Anal.*, **41** (2021), 84–121. https://doi.org/10.1093/imanum/drz074

18. C. W. Royer, M. O'Neill, S. J. Wright, A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization, *Math. Program.*, **180** (2020), 451–488. https://doi.org/10.1007/s10107-019-01362-7

19. P. Dvurechensky, M. Staudigl, Barrier algorithms for constrained non-convex optimization, preprint, arXiv:2404.18724. https://doi.org/10.48550/arXiv.2404.18724

20. C. He, Z. Lu, A Newton-CG based barrier method for finding a second-order stationary point of nonconvex conic optimization with complexity guarantees, *SIAM J. Optim.*, **33** (2023), 1191–1222. https://doi.org/10.1137/21m1457011

21. Y. Nesterov, A. Nemirovskii, *Interior-point Polynomial Algorithms in Convex Programming*, SIAM, 1994. https://doi.org/10.1137/1.9781611970791

22. Y. Nesterov, B. T. Polyak, Cubic regularization of Newton method and its global performance, *Math. Program.*, **108** (2006), 177–205. https://doi.org/10.1007/s10107-006-0706-8

23. A. Griewank, The modification of Newton's method for unconstrained optimization by bounding cubic terms, *Technical Report NA/12*, (1981). https://doi.org/10.13140/RG.2.1.4097.2960

24. J. Lagarias, R. Vanderbei, II Dikin's convergence result for the affine scaling algorithm, *Contemp. Math*, **114** (1990), 109–119. https://doi.org/10.1090/conm/114/1097868

25. E. R. Barnes, A variation on Karmarkar's algorithm for solving linear programming problems, *Math. Program.*, **36** (1986), 174–182. https://doi.org/10.1007/bf02592024

26. R. J. Vanderbei, M. S. Meketon, B. A. Freedman, A modification of Karmarkar's linear programming algorithm, *Algorithmica*, **1** (1986), 395–407. https://doi.org/10.1007/BF01840454

27. MATLAB Optimization Toolbox, MATLAB Optimization Toolbox, R2022a. The MathWorks, Natick, MA, USA. Available from: https://www.mathworks.com .

28. E. B. Gunnarsson, J. Foo, K. Leder, Statistical inference of the rates of cell proliferation and phenotypic switching in cancer, *J. Theor. Biol.*, **568** (2023), 111497. https://doi.org/10.1016/j.jtbi.2023.111497

29. A. F. Villaverde, F. Fröhlich, D. Weindl, J. Hasenauer, J. R. Banga, Benchmarking optimization methods for parameter estimation in large kinetic models, *Bioinformatics*, **35** (2019), 830–838. https://doi.org/10.1093/bioinformatics/bty736

30. D. K. Agrawal, B. J. Smith, P. D. Sottile, G. Hripcsak, D. J. Albers, Quantifiable identification of flow-limited ventilator dyssynchrony with the deformed lung ventilator model, *Comput. Biol. Med.*, **173** (2024), 108349. https://doi.org/10.1016/j.compbiomed.2024.108349
31. A. Köhn-Luque, E. M. Myklebust, D. S. Tadele, M. Giliberto, L. Schmiester, J. Noory, et al., Phenotypic deconvolution in heterogeneous cancer cell populations using drug-screening data, *Cell Rep. Methods*, **3** (2023), 100417. https://doi.org/10.1016/j.crmeth.2023.100417
32. R. Gesztelyi, J. Zsuga, A. Kemeny-Beke, B. Varga, B. Juhasz, A. Tosaki, The Hill equation and the origin of quantitative pharmacology, *Arch. Hist. Exact Sci.*, **66** (2012), 427–438. https://doi.org/10.1007/s00407-012-0098-5
33. A. V. Hill, The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves, *J. Physiol.*, **40** (1910), iv–vii.
34. S. R. Gadagkar, G. B. Call, Computational tools for fitting the Hill equation to dose–response curves, *J. Pharmacol. Toxicol. Methods*, **71** (2015), 68–76. https://doi.org/10.1016/j.vascn.2014.08.006
35. W. Jin, S. W. McCue, M. J. Simpson, Extended logistic growth model for heterogeneous populations, *J. Theor. Biol.*, **445** (2018), 51–61. https://doi.org/10.1016/j.jtbi.2018.02.027
36. C. Cartis, N. I. M. Gould, P. L. Toint, Adaptive cubic regularisation methods for unconstrained optimization. Part I: Motivation, convergence and numerical results, *Math. Program.*, **127** (2011), 245–295. https://doi.org/10.1007/s10107-009-0286-5
37. A. R. Conn, N. I. M. Gould, P. L. Toint, Trust region methods, *Soc. Ind. Appl. Math.*, (2000). https://doi.org/10.1137/1.9780898719857
38. D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, *J. Comput. Phys.*, **22** (1976), 403–434. https://doi.org/10.1016/0021-9991(76)90041-3

Appendix

The supplemental materials included here are structured as follows: Appendix A.1 provides the proof of Theorem 2.2 which provides the convergence rate of CRNAS; Appendix A.2 describes how to solve the crucial subproblem in CRNAS; Appendix A.3 describes a first-order version of CRNAS called first-order affine scaling and provides a theoretical analysis of its performance; Appendix A.4 presents all the candidate algorithms that we analyzed and compared in our numerical case studies, along with the corresponding metrics used for comparison. Appendices A.5–A.7 provide additional details about of our numerical experiments; lastly, Appendix A.8 states the termination criteria utilized by all of the algorithms in our experiments.

A.1. Complexity analysis of CRNAS

To begin, we quote Lemma 3 in [20], which shall prove crucial in our analysis.

Lemma A.1. *Based on Assumption 1, the following inequalities hold for all* $x, y \in \text{int}(\mathcal{K})$*:*

$$\left\|\nabla L(y) - \nabla L(x) - \nabla^2 L(x)(y-x)\right\|_x^* \leq \frac{1}{2}\beta\|y-x\|_x^2,$$

$$L(y) \leq L(x) + \nabla L(x)^\top (y-x) + \frac{1}{2}(y-x)^\top \nabla^2 L(x)(y-x) + \frac{1}{6}\beta \|y-x\|_x^3.$$

Observe that the affine scaling algorithm is to minimize the majorization. We show that the function value decreases can be lower bounded by the norm of $\nabla L(\theta^{k+1})$ projected in the linear space in the following two lemmas.

Lemma A.2. *If $M \geq 2\beta$, then the following inequality holds:*

$$L(\theta^k) - L(\theta^{k+1}) \geq \frac{M}{12}\|\theta^{k+1} - \theta^k\|_{\theta^k}^3.$$

Proof. Since θ^{k+1} is the optimal solution of the subproblem, we have the following:

$$\begin{aligned} L(\theta^k) &\geq L(\theta^k) + \left\langle \nabla L(\theta^k), \theta^{k+1} - \theta^k \right\rangle + \frac{1}{2}\left\langle \nabla^2 L(\theta^k)(\theta^{k+1} - \theta^k), \theta^{k+1} - \theta^k \right\rangle + \frac{M}{6}\|\theta^{k+1} - \theta^k\|_{\theta^k}^3 \\ &\geq L(\theta^{k+1}) + \frac{M}{12}\|\theta^{k+1} - \theta^k\|_{\theta^k}^3, \end{aligned}$$

where the last step is according to Lemma A.1. □

Because of the linear constraint $A\theta = b$, we expect $(\nabla L(\theta^{k+1}))^\top d$ to converge to zero after affine scaling, where d is any given feasible direction.

Lemma A.3. *Assume that the constraint $\|\theta - \theta^k\|_{\theta^k} \leq 1 - \alpha$ in the subproblem is inactive, namely $\|\theta^{k+1} - \theta^k\|_{\theta^k} < 1 - \alpha$. Then, for any d, such as $Ad = 0$, we have the following:*

$$|(\nabla L(\theta^{k+1}))^\top d| \leq \frac{M+\beta}{2}\|\theta^{k+1} - \theta^k\|_{\theta^k}^2 \|d\|_{\theta^k}$$

Proof. Because the constraint $\|\theta - \theta^k\|_{\theta^k} \leq 1 - \alpha$ is inactive, any d that satisfies $Ad = 0$ must be a feasible direction. According to the optimality condition, we have

$$\left(\nabla L(\theta^k) + \nabla^2 L(\theta^k)(\theta^{k+1} - \theta^k)\right)^\top d + \frac{M}{2}\|\theta^{k+1} - \theta^k\|_{\theta^k}(\theta^{k+1} - \theta^k)^\top (\nabla^2 B(\theta^k))d = 0,$$

which implies that

$$\begin{aligned} \left|\left(\nabla L(\theta^k) + \nabla^2 L(\theta^k)(\theta^{k+1} - \theta^k)\right)^\top d\right| &= \frac{M}{2}\|\theta^{k+1} - \theta^k\|_{\theta^k}|(\theta^{k+1} - \theta^k)^\top (\nabla^2 B(\theta^k))d| \\ &\leq \frac{M}{2}\|\theta^{k+1} - \theta^k\|_{\theta^k}^2 \|d\|_{\theta^k} \end{aligned}$$

According to Lemma A.1, we have $\left\|\nabla L(\theta^{k+1}) - \nabla L(\theta^k) - \nabla^2 L(\theta^k)(\theta^{k+1} - \theta^k)\right\|_{\theta^k}^* \leq \frac{\beta}{2}\|\theta^{k+1} - \theta^k\|_{\theta^k}^2$. Therefore,

$$\left|\left(\nabla L(\theta^{k+1}) - \nabla L(\theta^k) - \nabla^2 L(\theta^k)(\theta^{k+1} - \theta^k)\right)^\top d\right| \leq \frac{\beta}{2}\|\theta^{k+1} - \theta^k\|_{\theta^k}^2 \|d\|_{\theta^k}$$

Combining the above two, we get the inequality. □

We give the lemma below to show the second-order optimality of the solution obtained from the algorithm.

Lemma A.4. *Assume that the constraint $||\theta - \theta^k||_{\theta^k} \leq 1 - \alpha$ in the subproblem is inactive, namely $||\theta^{k+1} - \theta^k||_{\theta^k} < 1 - \alpha$. Then, for any d, such as $Ad = 0$, we have the following:*

$$d^\top \nabla^2 L(\theta^{k+1})d \geq -(M+\beta)||\theta^{k+1} - \theta^k||_{\theta^k} d^\top \nabla^2 B(\theta^k)d$$

Proof. Because the constraint $||\theta - \theta^k||_{\theta^k} \leq 1 - \alpha$ is inactive, any d that satisfies $Ad = 0$ must be a feasible direction. According to the second-order optimality condition, we have the following:

$$d^\top(\nabla^2 L(\theta^k) + \frac{M}{2}\frac{(\nabla^2 B(\theta^k))(\theta^{k+1} - \theta^k)(\theta^{k+1} - \theta^k)^\top(\nabla^2 B(\theta^k))}{||\theta^{k+1} - \theta^k||_{\theta^k}} + \frac{M}{4}||\theta^{k+1} - \theta^k||_{\theta^k}\nabla^2 B(\theta^k))d \geq 0$$

Therefore, we observe that

$$\begin{aligned} & d^\top\left(\nabla^2 L(\theta^k) + M||\theta^{k+1} - \theta^k||_{\theta^k}\nabla^2 B(\theta^k)\right)d \qquad (A.1)\\ \geq\ & d^\top\left(\nabla^2 L(\theta^k) + \frac{M}{2}(\nabla^2 B(\theta^k))(\theta^{k+1} - \theta^k)(\theta^{k+1} - \theta^k)^\top(\nabla^2 B(\theta^k))/||\theta^{k+1} - \theta^k||_{\theta^k}\right.\\ & \left. + \frac{M}{4}||\theta^{k+1} - \theta^k||_{\theta^k}\nabla^2 B(\theta^k)\right)d\\ \geq\ & 0. \end{aligned}$$

According to Assumption 1, we have the following:

$$\nabla^2 L(\theta^{k+1}) \geq \nabla^2 L(\theta^k) - \beta||\theta^{k+1} - \theta^k||_{\theta^k}\nabla^2 B(\theta^k).$$

Combining the above two inequalities, we have the following:

$$d^\top \nabla^2 L(\theta^{k+1})d \geq -(M+\beta)||\theta^{k+1} - \theta^k||_{\theta^k} d^\top \nabla^2 B(\theta^k)d.$$

□

According to Lemma A.2, we have

$$L(\theta^k) - L(\theta^{k+1}) \geq \frac{M}{12}||\theta^{k+1} - \theta^k||^3_{\theta^k} \geq \frac{M\eta^3}{12}, k = 0, 1, 2, ..., K-2,$$

which implies the decrease guarantee in each step so that the algorithm will stop in finite steps. By telescoping the above inequalities, we have the following:

$$\frac{M\eta^3}{12}(K-1) \leq \sum_{k=0}^{K-2}(L(\theta^k) - L(\theta^{k+1})) = L(\theta^0) - L(\theta^{K-1}) \leq L(\theta^0) - L^*.$$

Therefore, we can give an upper bound for the number of iterations $K \leq 12(L(\theta^0) - L^*)\eta^{-3} + 1$. According to Lemmas A.3 and A.4 with $M \geq 2\beta$ and $\eta \leq 1 - \alpha$, we have the following bounds.

For any d, such as $Ad = 0$, we have

$$|(\nabla L(\theta^K))^\top d| \leq \frac{M+\beta}{2}||\theta^K - \theta^{K-1}||^2_{\theta^{K-1}}||d||_{\theta^{K-1}} \leq M\eta^2||d||_{\theta^{K-1}}$$

and

$$d^\top \nabla^2 L(\theta^K)d \geq -(M+\beta)||\theta^K - \theta^{K-1}||_{\theta^{K-1}} d^\top \nabla^2 B(\theta^{K-1})d \geq -2M\eta d^\top \nabla^2 B(\theta^{K-1})d$$

Given that the right hand sides of the above two inequalities also involve θ^{K-1}, we cite the following lemma to change them into θ^K.

Lemma A.5. *(Theorem 2.1.1 in [21]) Let B be self-concordant on $\mathcal{K}$, and let $\theta_0 \in \text{int}(\mathcal{K})$; then, for any $\theta \in \text{int}(\mathcal{K})$ and $\|\theta - \theta_0\|_{\theta_0} < 1$, it holds that $(1 - \|\theta - \theta_0\|_{\theta_0})^2 u^\top \nabla^2 B(\theta_0) u \le u^\top \nabla^2 B(\theta) u \le \frac{u^\top \nabla B(\theta_0) u}{(1-\|\theta-\theta_0\|_{\theta_0})^2}$.*

Since $\|\theta^K - \theta^{K-1}\|_{\theta^{K-1}} < \eta \le 1 - \alpha$, we have $\alpha^2 \nabla^2 B(\theta^K) \le \nabla^2 B(\theta^{K-1}) \le \frac{1}{\alpha^2} \nabla^2 B(\theta^K)$. The above two inequalities imply

$$|(\nabla L(\theta^K))^\top d| \le M\eta^2 \|d\|_{\theta^{K-1}} \le M\eta^2 \alpha^{-1} \|d\|_{\theta^K}$$

and

$$d^\top \nabla^2 L(\theta^K) d \ge -2M\eta d^\top \nabla^2 B(\theta^{K-1}) d \ge -2M\eta\alpha^{-2} d^\top \nabla^2 B(\theta^K) d.$$

Then, Theorem 2.2 directly follows.

A.2. Solving the subproblem (2.7)

In each step, we solve the following subproblem:

$$\theta^{k+1} = \underset{\theta: A\theta = b, \|\theta-\theta^k\|_{\theta^k} \le 1-\alpha}{\arg\min} \left(\left\langle \nabla L\left(\theta^k\right), \theta - \theta^k \right\rangle + \frac{1}{2} \nabla^2 L\left(\theta^k\right) [\theta - \theta^k]^2 + \frac{M}{6} \left\| \theta - \theta^k \right\|_{\theta^k}^3 \right).$$

First, we eliminate the linear equation constraint $A\theta = b$ and replace the local norm with ℓ_2 norm by the following procedure. Let T be an orthogonal matrix whose columns form a basis of the linear space $A\theta = 0$. Then, the linear constraint $A\theta = b$ can be replaced by $\theta = \theta^k + T\theta'$. Let $\left\| \theta - \theta^k \right\|_{\theta^k} = \|(\nabla^2 B(\theta^k))^{1/2}(\theta - \theta^k)\| = \|(\nabla^2 B(\theta^k))^{1/2} T\theta'\| = \|(T^\top \nabla^2 B(\theta^k) T)^{1/2} \theta'\|$. The matrix $T^\top \nabla^2 B(\theta^k) T$ is invertible because $T^\top \nabla^2 B(\theta^k) T = ((\nabla^2 B(\theta^k))^{1/2} T)^\top ((\nabla^2 B(\theta^k))^{1/2} T)$, where T is of full column rank and $\nabla^2 B(\theta^k)$ is of full rank. Then, by letting $\bar{\theta} = (T^\top \nabla^2 B(\theta^k) T)^{1/2} \theta'$, the subproblem can be written as follows:

$$\min_{\|\bar{\theta}\| \le 1-\alpha} m(\bar{\theta}) := g^\top \bar{\theta} + \frac{1}{2} \bar{\theta}^\top P \bar{\theta} + \frac{M}{6} \|\bar{\theta}\|^3,$$

where g is a vector, and P is a matrix with corresponding sizes.

Then, we will show that the above problem is actually the standard cubic Newton subproblem or the trust-region subproblem, either of which can be easily solved by well-known methods. Note in the subproblem of standard cubic Newton, we solve $\min_{\bar{\theta}} m(\bar{\theta})$, where we confine $\|\bar{\theta}\| \le 1 - \alpha$ in the above subproblem. If the solution to $\min_{\bar{\theta}} m(\bar{\theta})$ satisfies $\|\bar{\theta}^*\| \le 1 - \alpha$, then we get the solution. Otherwise, we will show that it is actually a trust-region subproblem from the following two lemmas. We introduce the following two lemmas to solve the subproblem.

Lemma A.6. *Let $\hat{\theta} \in \arg\min_{\|\bar{\theta}\| \le 1-\alpha} m(\bar{\theta})$ and $\bar{\theta}^* \in \arg\min_{\bar{\theta}} m(\bar{\theta})$. If $\|\bar{\theta}^*\| > 1 - \alpha$, then we must have $\|\hat{\theta}\| = 1 - \alpha$.*

Proof. We cited part of the proof of Theorem 3.1 in [36]. Suppose $\|\hat{\theta}\| < 1 - \alpha$; we will show that $\hat{\theta}$ is a global minimum and no global solution will satisfy $\|\theta\| < 1 - \alpha$. Since $\hat{\theta}$ is an interior point of the set $\{\theta : \|\theta\| \le 1 - \alpha\}$, according to the first and second-order necessary optimality conditions,

$$g + (P + \hat{\lambda} I)\hat{\theta} = \mathbf{0}$$

and

$$w^\top \left(P + \hat{\lambda} I + \hat{\lambda} \left(\frac{\hat{\theta}}{\|\hat{\theta}\|} \right) \left(\frac{\hat{\theta}}{\|\hat{\theta}\|} \right)^\top \right) w \ge 0 \tag{A.2}$$

for all vectors w, where $\hat{\lambda} = \frac{M}{2}\|\hat{\theta}\|$. If $\|\hat{\theta}\| = 0$, the the second-order optimizality condition is $P + \hat{\lambda} I \geq 0$. We will show that $P + \hat{\lambda} I \geq 0$ still holds for $\|\hat{\theta}\| \neq 0$.

For $w^\top \hat{\theta} = 0$, (A.2) shows that $w^\top (P + \hat{\lambda} I) w \geq 0$. Consider the vector $w^\top \hat{\theta} \neq 0$; the line $\hat{\theta} + \alpha w$ intersects the ball of radius $\hat{\theta}$ at two points, $\hat{\theta} \neq \theta'$, where $\|\hat{\theta}\| = \|\theta'\|$. Without loss of generality, let $w = \theta' - \hat{\theta}$. Since $\hat{\theta} \in \arg\min_{\|\bar{\theta}\| \leq 1-\alpha} m(\bar{\theta})$, we have the following:

$$\begin{aligned}
0 &\leq m(\theta') - m(\hat{\theta}) \\
&= g^\top(\theta' - \hat{\theta}) + \frac{1}{2}\theta'^\top P \theta' - \frac{1}{2}\hat{\theta}^\top P \hat{\theta} \\
&= \hat{\theta}^\top (P + \hat{\lambda} I)(\hat{\theta} - \theta') + \frac{1}{2}\theta'^\top P \theta' - \frac{1}{2}\hat{\theta}^\top P \hat{\theta} \\
&= \hat{\lambda}\|\hat{\theta}\|^2 - \hat{\lambda}\hat{\theta}^\top \theta' - \hat{\theta}^\top P \theta' + \frac{1}{2}\theta'^\top P \theta' + \frac{1}{2}\hat{\theta}^\top P \hat{\theta} \\
&= \frac{1}{2}\hat{\lambda}\|\hat{\theta}\|^2 + \frac{1}{2}\hat{\lambda}\|\theta'\|^2 - \hat{\lambda}\hat{\theta}^\top \theta' - \hat{\theta}^\top P \theta' + \frac{1}{2}\theta'^\top P \theta' + \frac{1}{2}\hat{\theta}^\top P \hat{\theta} \\
&= \frac{1}{2}\hat{\lambda}(\theta' - \hat{\theta})^\top (\theta' - \hat{\theta}) + \frac{1}{2}(\theta' - \hat{\theta})^\top P (\theta' - \hat{\theta}) \\
&= \frac{1}{2} w^\top (P + \hat{\lambda} I) w
\end{aligned}$$

Therefore, $P + \hat{\lambda} I \geq 0$ always holds. Let P have an eigendecomposition $P = U^\top \Lambda U$, where $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_n)$. Then, according to A.2, we have the following:

$$Ug + (\Lambda + \hat{\lambda} I) U \hat{\theta} = \mathbf{0},$$

where $\hat{\lambda} = \frac{M}{2}\|\hat{\theta}\|$, and $\lambda_i + \hat{\lambda} \geq 0$, $i = 1, 2, ..., n$.

On the other hand, according to the first-order optimality condition for the global solution $\bar{\theta}^*$, we have

$$g + (P + \bar{\lambda}^* I)\bar{\theta}^* = \mathbf{0},$$

which is equivalent to

$$Ug + (\Lambda + \bar{\lambda}^* I) U \bar{\theta}^* = \mathbf{0},$$

where $\bar{\lambda}^* = \frac{M}{2}\|\bar{\theta}^*\| \geq \frac{M}{2}(1 - \alpha) > \hat{\lambda}$. Therefore, we have $\lambda_i + \bar{\lambda}^* > \lambda_i + \hat{\lambda} \geq 0$, $i = 1, 2, ..., n$. Thus, we can get the following equation:

$$U\bar{\theta}^* = (\Lambda + \bar{\lambda}^* I)^{-1}(-Ug) = (\Lambda + \bar{\lambda}^* I)^{-1}(\Lambda + \hat{\lambda} I) U \hat{\theta},$$

which implies $|(U\bar{\theta}^*)_i| = |\frac{\lambda_i + \hat{\lambda}}{\lambda_i + \bar{\lambda}^*}(U\hat{\theta})_i| \leq |(U\hat{\theta})_i|$. Therefore, we have $\|U\bar{\theta}^*\| \leq \|U\hat{\theta}\|$. However, since U is orthogonal,

$$\|U\bar{\theta}^*\| = \|\bar{\theta}^*\| \geq 1 - \alpha > \|\hat{\theta}\| = \|U\hat{\theta}\|,$$

which causes a contradiction. □

Lemma A.7. *Let* $\tilde{\theta} \in \arg\min_{\|\bar{\theta}\| \leq 1-\alpha} g^\top \bar{\theta} + \frac{1}{2}\bar{\theta}^\top P \bar{\theta}$ *and* $\bar{\theta}^* \in \arg\min_{\bar{\theta}} m(\bar{\theta})$. *If* $\|\bar{\theta}^*\| > 1 - \alpha$, *then*

$$\tilde{\theta} \in \arg\min_{\|\bar{\theta}\| \leq 1-\alpha} m(\bar{\theta})$$

Proof. Suppose $\tilde{\theta} \notin \arg\min_{\|\bar{\theta}\|\leq 1-\alpha} m(\bar{\theta})$. Let $\hat{\theta} \in \arg\min_{\|\bar{\theta}\|\leq 1-\alpha} m(\bar{\theta})$. Then, we have $m(\tilde{\theta}) > m(\hat{\theta})$, which is

$$g^\top\tilde{\theta} + \frac{1}{2}\tilde{\theta}^\top P\tilde{\theta} + \frac{M}{6}\|\tilde{\theta}\|^3 > g^\top\hat{\theta} + \frac{1}{2}\hat{\theta}^\top P\hat{\theta} + \frac{M}{6}\|\hat{\theta}\|^3$$

According to Lemma A.6, we have $\|\hat{\theta}\| = 1-\alpha \geq \|\tilde{\theta}\|$. Therefore,

$$g^\top\tilde{\theta} + \frac{1}{2}\tilde{\theta}^\top P\tilde{\theta} > g^\top\hat{\theta} + \frac{1}{2}\hat{\theta}^\top P\hat{\theta} + \frac{M}{6}(\|\hat{\theta}\|^3 - \|\tilde{\theta}\|^3) \geq g^\top\hat{\theta} + \frac{1}{2}\hat{\theta}^\top P\hat{\theta},$$

which contradicts with the fact $\tilde{\theta} \in \arg\min_{\|\bar{\theta}\|\leq 1-\alpha} g^\top\bar{\theta} + \frac{1}{2}\bar{\theta}^\top P\bar{\theta}$. □

Therefore, when solving the subproblem, we first solve the problem $\min_{\bar{\theta}} m(\bar{\theta})$ by the method introduced in Section 6.1 of [36]. If $\|\bar{\theta}^*\| \leq 1-\alpha$, then we find the solution $\bar{\theta}^*$. If not, then we know the optimal solution is on the boundary, namely $\|\hat{\theta}\| = 1-\alpha$. Thus, we can skip the cubic term by adding the ball constraints. Only the following trust-region subproblem needs to be solved:

$$\hat{\theta} = \arg\min_{\|\bar{\theta}\|\leq 1-\alpha} g^\top\bar{\theta} + \frac{1}{2}\bar{\theta}^\top P\bar{\theta},$$

which is discussed in 7.3 and 1.3 of [37].

A.3. A first-order version of CRNAS

Here, we discuss a first-order version of CRNAS to solve (2.2). The construction of our first-order method is essentially the same as that for CRNAS, with the exception that a quadratic approximation of the function is utilized rather than a cubic approximation. Hence, the subproblem solved at each iteration of our first-order affine scaling method (FOAS) is as follows:

$$\theta^{k+1} = \arg\min_{\theta: A\theta=b, \|\theta-\theta^k\|_{\theta^k}\leq 1-\alpha} \left(\left\langle \nabla L\left(\theta^k\right), \theta-\theta^k\right\rangle + \frac{M}{2}\left\|\theta-\theta^k\right\|_{\theta^k}^2\right),$$

where M is a positive number. The exact description of the algorithm is provided below; you will recognize it as identical to CRNAS modulo the altered subproblem.

First-order affine scaling

Step 0: Provide an interior point θ^0 (i.e., $A\theta^0 = b$ and $\theta^0 \in \text{int}(\mathcal{K})$); choose the constants $\eta > 0$, $M > 0$, and $\alpha \in (0,1)$; set $k = 0$

Step 1: For $k = 0, 1, ..., K-1$, solve the following subproblem:

$$\theta^{k+1} = \arg\min_{\theta: A\theta=b, \|\theta-\theta^k\|_{\theta^k}\leq 1-\alpha} \left(\left\langle \nabla L\left(\theta^k\right), \theta-\theta^k\right\rangle + \frac{M}{2}\left\|\theta-\theta^k\right\|_{\theta^k}^2\right)$$

Step 2: If $\|\theta^{k+1}-\theta^k\|_{\theta^k} < \eta$, then let $K = k+1$ and stop. Otherwise, go back to Step 1 with $k = k+1$

As CRNAS was well-defined, so is FOAS due to Lemma 2.1, which guarantees all of the generated iterates lie inside the interior of the cone. Therefore, we proceed to prove the convergence rate of FOAS. To begin, we assume the Lipschitz smoothness of the gradient.

Assumption 2. *There exists a constant $\beta \geq 0$ such that for all $x, y \in \text{int}(\mathcal{K})$,*

$$\|\nabla L(y) - \nabla L(x)\|_x^* \leq \beta \|y - x\|_x.$$

Using this assumption, it follows that there exists a local quadratic upper bound for L at all points inside the interior of the cone.

Lemma A.8. *Under Assumption 2, the following inequality holds for all $x, y \in \text{int}(\mathcal{K})$:*

$$L(y) \leq L(x) + \nabla L(x)^\top (y - x) + \frac{\beta}{2}\|y - x\|_x^2.$$

Proof. According to the fundamental theorem of calculus and norm relations, we have the following:

$$\begin{aligned} L(y) - L(x) - \nabla L(x)^\top (y - x) &= \left\langle \int_0^1 (\nabla L(x + t(y - x)) - \nabla L(x)) dt, y - x \right\rangle \\ &\leq \left\| \int_0^1 (\nabla L(x + t(y - x)) - \nabla L(x)) dt \right\|_x^* \|y - x\|_x \\ &\leq \left(\int_0^1 \|(\nabla L(x + t(y - x)) - \nabla L(x))\|_x^* \, dt \right) \|y - x\|_x \\ &\leq \left(\int_0^1 \beta \|t(y - x)\|_x \, dt \right) \|y - x\|_x \\ &= \frac{\beta}{2}\|y - x\|_x^2 \end{aligned}$$

where the last inequality follows from Assumption 2. □

Lemma A.8 demonstrates that FOAS at each iteration minimizes a local quadratic upper bound of our objective function, provided M is large enough. We show that the function value decreases can be lowered bounded by the norm of $\nabla L(\theta^{k+1})$ projected in the linear space in the following two lemmas.

Lemma A.9. *If $M \geq 2\beta$, then for all the iterates generated by FOAS, we have that*

$$L(\theta^k) - L(\theta^{k+1}) \geq \frac{M}{4}\|\theta^{k+1} - \theta^k\|_{\theta^k}^2.$$

Proof. Since θ^{k+1} is the optimal solution of the FOAS subproblem, we have the following:

$$\begin{aligned} L(\theta^k) &\geq L(\theta^k) + \left\langle \nabla L(\theta^k), \theta^{k+1} - \theta^k \right\rangle + \frac{M}{2}\|\theta^{k+1} - \theta^k\|_{\theta^k}^2 \\ &\geq L(\theta^{k+1}) + \frac{M}{4}\|\theta^{k+1} - \theta^k\|_{\theta^k}^2, \end{aligned}$$

where the last step follows from Lemma A.8. □

Because of the linear constraint $A\theta = b$, we expect $(\nabla L(\theta^k))^\top d$ to converge to zero after affine scaling, where d is any given feasible direction.

Lemma A.10. *Assume that the constraint $||P\theta - \theta^k||_{\theta^k} \leq 1 - \alpha$ in the subproblem is inactive, namely $||\theta^{k+1} - \theta^k||_{\theta^k} < 1 - \alpha$. Then, for any d such as $Ad = 0$, we have the following:*

$$|\left(\nabla L(\theta^k)\right)^\top d| \leq M||\theta^{k+1} - \theta^k||_{\theta^k}||d||_{\theta^k}.$$

Proof. Because the constraint $||\theta - \theta^k||_{\theta^k} \leq 1 - \alpha$ is inactive, any d that satisfies $Ad = 0$ must be a feasible direction. According to the optimality condition, we have the following:

$$\left(\nabla L(\theta^k) + M(\nabla^2 B(\theta^k))(\theta^{k+1} - \theta^k)\right)^\top d = 0,$$

which implies

$$|\left(\nabla L(\theta^k)\right)^\top d| = M|(\theta^{k+1} - \theta^k)^\top(\nabla^2 B(\theta^k))d| \leq M||\theta^{k+1} - \theta^k||_{\theta^k}||d||_{\theta^k}.$$

□

Given the fact $\sum_{k=0}^{K-1}(L(\theta^k) - L(\theta^{k+1})) = L(\theta_0) - L(\theta^K) \leq L(\theta_0) - L^*$, there exists a $0 \leq k \leq K-1$, such that $L(\theta^k) - L(\theta^{k+1}) \leq \frac{L(\theta_0)-L^*}{K}$.

Therefore, with Lemma A.9, we obtain

$$\frac{L(\theta^0) - L^*}{K} \geq L(\theta^k) - L(\theta^{k+1}) \geq \frac{M}{4}||\theta^{k+1} - \theta^k||^2_{\theta^k},$$

which implies

$$||\theta^{k+1} - \theta^k||^2_{\theta^k} \leq \frac{4(L(\theta^0) - L^*)}{KM}.$$

Therefore, as long as $K > \frac{4(L(\theta^0)-L^*)}{M(1-\alpha)^2}$ and $M \geq 2\beta$, according to Lemma A.10, we have the following bounds.

For any d, such as $Ad = 0$, we have

$$|(\nabla L(\theta^k))^\top d| \leq M||\theta^{k+1} - \theta^k||_{\theta^k}||d||_{\theta^k} \leq 2(L(\theta^0) - L^*)^{1/2}M^{1/2}K^{-1/2}||d||_{\theta^k}.$$

We arrive at the following theorem.

Theorem A.11. *Use First-Order Affine Scaling Algorithm A.3 and assume Assumption 2 and $M > 2\beta$. For any $K > \frac{4(L(\theta^0)-L^*)}{M(1-\alpha)^2}$, there is a $0 \leq k \leq K-1$, such that for any d with $Ad = 0$, we have the following:*

$$|(\nabla L(\theta^k))^\top d| \leq 2(L(\theta^0) - L^*)^{1/2}M^{1/2}K^{-1/2}||d||_{\theta^k}.$$

A.4. Benchmarking algorithms and performance metrics

Since CRNAS relies on a precise gradient and Hessian information, we considered implementations of algorithms in *fmincon* both with and without specifying the gradient and Hessian of the objective function. We implemented both the *interior-point* and *sequential quadratic programming* algorithms in *fmincon* by only supplying the objective function and allowing MATLAB's built-in techniques to estimate the gradient and Hessian; IP and SQP refer to these implementations. Additionally, we tested the *interior-point* solver with the exact gradient and Hessian; IP hess denotes this instantiation. Lastly,

Table A1. Summary of the algorithms tested in our numerical experiments.

Algorithm	Gradient/Hessian	Acronym
Cubic regularized Newton with affine scaling	Yes	CRNAS
Interior-point in *fmincon*	No	IP
Interior-point in *fmincon*	Yes	IP hess
Sequential quadratic programming in *fmincon*	No	SQP
Sequential quadratic programming in *fmincon*	Yes	SQP grad

the sequential quadratic programming method with the exact gradient was employed in our testing; SQP grad denotes this procedure. Table A1 summarizes these different approaches.

Our studies sought to analyze the effectiveness and quality of the solutions provided by each algorithm. We evaluated three quantities in our assessment of each method:

1) **Total compute time**—the total wall time required for Algorithm x to run from N different initial points and terminate. Termination will occur if a local solution is obtained up to some provided tolerance or some computational resource or numerical limit has been expended or reached (e.g., maximum iterations or minimum stepsize, respectively)
2) **Best computed value**—is the lowest value of the objective function obtained from Algorithm x from N different initializations; this value is Algorithm x's best estimate of the global minimum after N different instantiations
3) **Number of iterations to best value**—is the number of iterations required to obtain the best computed value from the single initial point which generated it

The first two metrics seek to provide a measure of which method shall be of the greatest benefit to a practitioner. These metrics measure which algorithm is most effective at producing the best solution the quickest when a multi-start approach is applied to estimate a global minimum. On the other hand, the third metric provides an estimate of total iterations needed for the method to compute its best solution from a single initialization. To ensure a fair comparison, in all of our tests, each algorithm was provided the same initial guesses and had the same termination criteria. The termination criteria used in the numerical experiments are detailed in Appendix A.8. In the next two sections, we set-up parameter estimation problems that arise in mathematical biology and compare CRNAS against *fmincon* using these three measures.

A.5. Deterministic drug-affected cell proliferation experiment details

To set up our optimization problems, we collected *in silico* data according to a true parameter set $\theta^*_{PP}(S)$, the deterministic framework at collections of time points $\mathcal{T}$, and drug dose levels $\mathcal{D}$. Without further specification, we employ the following time points and drug dose levels:

$$\begin{aligned} \mathcal{T} &= \{0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36\} \\ \mathcal{D} &= \{0, 0.0313, 0.0625, 0.125, 0.25, 0.375, 0.5, 1.25, 2.5, 3.75, 5\}. \end{aligned} \tag{A.3}$$

These settings are similar to those used in [31], which analyzes real-world *in vitro* datasets. We selected the true parameter set $\theta^*_{PP}(S)$ randomly from a biologically feasible range, denoted as $\Theta^*(S)$. This range

is specified for each experiment in the subsequent sections. With the true parameter set, we obtained the cell count data of each sub-type and the following total cell count data according to (3.1).

Based on the data, we solved (3.3) to obtain the point estimation $\hat{\theta}_{PP}(S)$ within the optimization feasible range $\hat{\Theta}(S)$. The optimization feasible range is larger than the biologically feasible region, thus indicating incomplete prior knowledge about the "true nature" of the parameters.

Experiment with $S = 1$

The biologically feasible range and optimization feasible range for this experiment are described in Table A2. Note that the true EC50 parameter E should be located within the drug dose levels $\mathcal{D}$ to ensure parameter identifiability.

Table A2. Biologically feasible range and optimization feasible range for the deterministic framework experiment with $S = 1$.

	α	b	E	n
$\Theta^*(1)$	$(0, 0.1)$	$(0.8, 1)$	$(0.05, 0.1)$	$(1.5, 5)$
$\hat{\Theta}(1)$	$(0, 1)$	$(0, 1)$	$(0, \infty)$	$(0, \infty)$

Experiment with $S = 2$

When $S = 2$, we denote the subpopulation with smaller EC50 value as "sensitive" and the subpopulation with a higher EC50 value as "resistant". The sensitive subpopulation EC50 parameter E_s is selected to be distinct from E_r of the resistant subpopulation for model identifiability. Other than the EC50 values, we do not distinguish the parameter space between the sensitive and resistant subpopulations. Additionally, we note that the initial proportions p_s and p_r are selected to satisfy $p_s + p_r = 1$. Table A3 outlines the biologically feasible range and the feasible region used in the optimization model for this experiment.

Table A3. Biologically feasible range and optimization feasible range for the deterministic framework experiment with $S = 2$.

	p	α	b	E_s	E_r	n
$\Theta^*(2)$	$(0, 1)$	$(0, 0.1)$	$(0.8, 1)$	$(0.05, 0.1)$	$(0.5, 2.5)$	$(1.5, 5)$
$\hat{\Theta}(2)$	$(0, 1)$	$(0, 1)$	$(0, 1)$	$(0, \infty)$	$(0, \infty)$	$(0, \infty)$

Experiment with $S > 2$

In this experiment, we dynamically selected the drug dose levels. Specifically, we selected the EC50 biologically feasible range and dose levels based on the number of subpopulations S. We kept the overall range of $\mathcal{D}$ the same (i.e., $\min(\mathcal{D}) = 0, \max(\mathcal{D}) = 10$), and selected $4S$ dose levels within this range according to a logarithmic scale. For example, with $S = 3$, the drug doses could be selected as follows:

$$\mathcal{D}(3) = \{0, 0.01, 0.02, 0.0398, 0.0794, 0.1585, 0.3162, 0.631, 1.2589, 2.5119, 5.0119, 10\}.$$

To preserve the parameter identifiability, we designed the biologically feasible ranges for the EC50 parameters based on the generated dose levels. For instance, we have the following biologically feasible ranges for the experiment $S = 3$ described in Table A4, where E_1, E_2, and E_3 are from distinct subpopulations and the ranges of the other parameters are the same across the subpopulations.

Table A4. Biologically feasible range and optimization feasible range for the deterministic framework experiment with $S = 3$.

	p	α	b	E_1	E_2	E_3	n
$\Theta^*(3)$	$(0, 1)$	$(0, 0.1)$	$(0.8, 1)$	$(0.005, 0.0299)$	$(0.119, 0.4736)$	$(1.8854, 7.5059)$	$(1.5, 5)$
$\hat{\Theta}(3)$	$(0, 1)$	$(0, 1)$	$(0, 1)$	$(0, \infty)$	$(0, \infty)$	$(0, \infty)$	$(0, \infty)$

Experiment using real-world data (*in vitro*)

In fitting the *in vitro* dataset in [31], we primarily rely on the optimization feasible ranges defined in Table A3. Unlike the noise-free model, we additionally introduce a noise parameter σ to represent the standard deviation of the observation noise. This parameter is estimated within the range $0 \leq \sigma \leq 1000$.

A.6. Stochastic drug-affected cell proliferation experiment details

Similar to the deterministic framework experiment, we randomly generated 100 true parameter sets $\theta^*_{LBD}(S)$ from biologically feasible ranges and the corresponding simulated dataset. Different from the PhenoPop experiment, we assumed that the stochastic linear birth-death process governed the tumor dynamic. Consequently, we employed the Gillespie algorithm [38] to simulate the data according to a stochastic process. Due to the stochastic nature of the simulation, we collected 13 replicates under the same drug dose levels and time points as in (A.3). The biologically feasible ranges and the constraints for the optimization problems are presented in Table A5.

Table A5. Biologically feasible range and optimization feasible range for the stochastic framework experiment with $S = 2$.

	p	β	ν	b	E_s	E_r	n
Θ^*	$(0, 1)$	$(\nu, \nu + 0.1)$	$(0, 1)$	$(0.8, 1)$	$(0.05, 0.1)$	$(0.5, 2.5)$	$(1.5, 5)$
$\hat{\Theta}$	$(0, 1)$	$(0, 1)$	$(0, 1)$	$(0, 1)$	$(0, \infty)$	$(0, \infty)$	$(0, \infty)$

A.7. Heterogeneous logistic model experiment details

Similar to the experiments in Section 3.1, we examined the algorithms' performance for 100 independent experiments with distinct true parameter sets $\theta^*_{LG}(S)$. The true parameter sets were randomly selected from the biologically feasible range shown in Table A6. With the true parameter set, we generated the experimental data at time points

$$\mathcal{T} = \{0, 1.111, 2.222, 3.333, 4.444, 5.555, 6.666, 7.777, 8.888, 10\},$$

according to the relation described in (3.6). Then, the optimization problems were formulated using the feasible regions described in Table A6 as in the prior experiments.

A.8. CRNAS stopping criteria

To ensure the fairness of comparison in the computational time of each optimization algorithm, we employed two stopping criteria for CRNAS, which were also implemented in fmincon. One was

Table A6. Biologically feasible range and optimization feasible range for the heterogeneous logistic growth model with $S = 2$.

	p_1	α_1	β_1	p_2	α_2	β_2
Θ^*	$(0, 1)$	$(0, 1)$	$(0, 1)$	$1 - p_1$	$(2, 3)$	$(2, 3)$
$\hat{\Theta}$	$(0, 1)$	$(0, 10)$	$(0, 10)$	$(0, 1)$	$(0, 10)$	$(0, 10)$

the first-order optimality condition, and the other was the stepsize of each iteration. In particular, the algorithm stopped once one of the following quantities falls below the given threshold $\epsilon = 10^{-6}$:

$$\begin{cases} \|\nabla f(\theta_{k+1})\| & \text{Euclidean norm of the gradient,} \\ \|\nabla f(\theta_{k+1})\|_{\theta_{k+1}} & \text{Local norm of the gradient,} \\ \|\theta_{k+1} - \theta_k\| & \text{Euclidean norm of the difference in each iteration solution,} \end{cases}$$

where $f(\theta)$ is the objective function. It is worth noting that *fmincon* does not use this measurement for the first-order optimality condition. However, both methods of computing the first-order optimality conditions were zero at the minimum. Therefore, we used the same threshold for both CRNAS and the *fmincon* algorithms. In addition to these two stopping criteria, the algorithm stopped if the number of iterations was greater than 500.

AIMS Press