TOWARDS LEARNING STOCHASTIC POPULATION MODELS BY GRADIENT DESCENT

PREPRINT

Justin N. Kreikemeyer[1], **Philipp Andelfinger**[1], **and Adelinde M. Uhrmacher**[1]

[1]Institute for Visual and Analytic Computing, University of Rostock; 18059 Rostock, Germany.
{justin.kreikemeyer, philipp.andelfinger, adelinde.uhrmacher}@uni-rostock.de

arXiv:2404.07049v1 [cs.LG] 10 Apr 2024

ABSTRACT

Increasing effort is put into the development of methods for learning mechanistic models from data. This task entails not only the accurate estimation of parameters, but also a suitable model structure. Recent work on the discovery of dynamical systems formulates this problem as a linear equation system. Here, we explore several simulation-based optimization approaches, which allow much greater freedom in the objective formulation and weaker conditions on the available data. We show that even for relatively small stochastic population models, simultaneous estimation of parameters and structure poses major challenges for optimization procedures. Particularly, we investigate the application of the local stochastic gradient descent method, commonly used for training machine learning models. We demonstrate accurate estimation of models but find that enforcing the inference of parsimonious, interpretable models drastically increases the difficulty. We give an outlook on how this challenge can be overcome.

Keywords automatic model generation, gradient descent, stochastic simulation algorithm, discrete-event simulation, differentiable simulation

1 Introduction

Statistical machine learning methods provide exciting advances in automatically learning (statistical) models from data. Whereas these models provide impressive predictive abilities [27], their black-box nature does not directly contribute to understanding the reference system's mechanics and impedes precise manual refinement. This motivated the development of methods for automatically deriving mechanistic models from data [25, 4, 15, 5]. With these, manual, hypothesis-driven knowledge discovery can increasingly be augmented by automatic, data-driven approaches [20]. Such an automatic modeling approach is useful when (parts of) the mechanisms of the reference system are unknown, but there are measurements of its behavior over time. Learning mechanistic models from data then entails not only parameter estimation but also the *simultaneous* identification of a suitable model structure.

In this paper, we study the case of learning stochastic, *discrete-event* models with an underlying continuous representation of time from *time-series* snapshots of some traversed state distributions by gradient descent. Specifically, we focus on Markovian *population models* that are expressed as reaction systems. Our contributions are as follows:

- Section 5 provides different possible formulations of the model learning problem.
- Section 5.1 shows how reparametrization enables parameter estimation over different orders of magnitude.
- Section 6 provides first results on the simultaneous learning of structure and parameters by gradient descent. It discusses the challenges and opportunities of the approach.

We briefly introduce the reaction system formalism in Section 2 and stochastic gradient estimation in Section 3. Section 4 reviews related work. After presenting our methods in Section 5 as outlined above, we conclude in Section 6.

2 Population-based Modeling

In the biology and chemistry domains, *reaction systems* are a commonly used modeling formalism [13]. They describe system dynamics in terms of the consumption and production of entities at certain rates. Their underlying assumption is that entities can be grouped into homogeneous populations of *species* (or molecules) $S_i, i \in \{1, \ldots, n_s\}$ that reside in a *well-stirred* medium. A reaction takes the form

$$R_i : \sum_{j=1}^{n_S} c_{ij} S_j \xrightarrow{r_i} \sum_{j=n_S+1}^{2n_S} c_{ij} S_{j-n_S}$$

with $\mathbf{C} \in \mathbb{N}^{n_R \times 2n_S}$ being a matrix of *coefficients* ("model structure"), $\mathbf{r}$ the vector of *rate constants* ("parameters"), and

© ACM 2024. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record will appear at the ACM SIGSIM PADS'24.

n_R, n_S the number of reactions and species in the system, respectively. A reaction system can be completely characterized by providing $\mathbf{C}$ and $\mathbf{r}$. A vector of species counts gives the starting conditions of a reaction system, i.e., $\mathbf{S}_{init}$.

As a running example, consider the well-studied SIR model of disease spread, comprising three species representing populations of **s**usceptible, **i**nfected and **r**ecovered individuals:

$$\begin{array}{ll} R_0\colon 1S + 1I & \xrightarrow{0.02} 2I \\ R_1\colon 1I & \xrightarrow{5.00} 1R \end{array} , \mathbf{C}_{sir} = \begin{pmatrix} 1 & 1 & 0 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

This reaction system has two reactions with coefficient matrix $\mathbf{C}_{sir}$. Here, the parameters $\mathbf{r}$ are chosen as $(0.02, 5)$. The first reaction describes the infection of a susceptible individual and the second an individual's recovery. Note that species participating with coefficient 0 are omitted. We will use $\mathbf{S}_{init} = (1980, 20, 0)$ as initial condition.

Population-based models defined as reaction systems can be simulated either through *numerical integration* with ordinary differential equation (ODE) semantics [18, 11] or the *stochastic simulation algorithm* (SSA) [9] with continuous-time Markov chain (CTMC) semantics. In many cases, stochastic effects cannot be ignored [30, 23]. Therefore, instead of focusing on the mean continuous dynamics, our approach will take the stochasticity of the system into account.

The vector of species counts $\mathbf{S}$ fully represents the state of the model at the current time t. We make the common assumption that the transition probabilities are governed by the probability of two entities in the well-stirred medium reacting and the transitions of the CTMC are governed by the stochastic mass action law [18, 9]. The effective rate of a reaction in a given state is called its *propensity* α. For example, for the SIR model, we have $\alpha_0 = 0.02 \cdot S \cdot I$, i.e., the more susceptible and infected individuals there are, the likelier an infection event is to happen. Note that other functions may be used to calculate the propensity depending on the modeled system. Another common assumption is that the probability of more than two species colliding is so low that any reaction with three or more reactants can be split into multiple reactions with two or fewer reactants. Thus, we only consider binary reactions with at most two species on the left-hand side of a reaction. Despite making these assumptions here for simplicity, our approach is theoretically able to accommodate any dependence of the propensities on the state as well as n-ary reactions.

As a simulator, we use Gillespie's direct method [9], which takes sample trajectories through the CTMC defined by $\mathbf{C}$ and $\mathbf{r}$ using a Monte Carlo strategy. At each event, t is advanced according to an exponential distribution over the sum of propensities α_i. The state is updated by choosing from a categorical distribution over the reactions, subtracting the reactants and adding the products to the current state. With the number of sample trajectories tending to infinity, the likelihood function of the model is recovered, i.e., the probability distribution over states and time given $\mathbf{r}$.

3 Stochastic Gradient Estimation

When there is a closed form of the likelihood, its gradient is an effective tool for optimization. However, the closed form is intractable for many real-world systems, necessitating Gillespie's SSA. Determining the gradient of this algorithm is not straightforward. The well-established method of automatic differentiation (AD) provides performant means to calculate the gradient of algorithms at runtime [21]. However, this gradient cannot account for the jumps (discontinuities) inherent to the individual SSA trajectories. So even with the mean over trajectories being a smooth function, AD is not useful for optimization.

Thus, we resort to recent advances in estimating the gradient of an alternative objective function, which is smoothed over jumps [17]. We use a finite-differences estimator with stochastic step-size, as proposed in [29] (Chapter 3.4) and further analyzed in [24]:

$$\nabla f(\theta) \approx \frac{1}{N} \sum_{n=1}^{N} \frac{f(\theta + \sigma \mathbf{u}) - f(\theta)}{\sigma} \mathbf{u} \tag{1}$$

where θ is the parameter vector, σ is a smoothing factor that determines the smoothing applied to the objective f and $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a vector of i.i.d. normal variates with mean 0 and variance 1. In contrast to finite differences, which need at least one sample per dimension of θ, through simultaneous perturbation this estimator requires only two samples for estimating the full gradient. For the number of samples n approaching infinity, the estimate converges to the gradient of a smoothed version of f [24]. Further, it can handle jumps and noise in the objective through the smoothing controlled by σ.

4 Related Work

The idea of learning mechanistic models or automating the modeling process has inspired various research in many application fields [16, 31, 3]. Related to our work, two major approaches can be distinguished: genetic programming, which for the first time provided strategic means of searching in the space of programs or models [16, 25], and sparse regression, which enables the identification of short yet accurate symbolic expressions, such as differential equations [8, 4]. Recently, these approaches have also been combined, e.g., to discover multibody physics systems [3].

Specifically in the case of biochemical reaction models, [25] proposed genetic programming to identify reaction systems with ODE semantics. Here, a population of candidate structures is evolved, and evolutionary operators are applied based on the candidates' fitness. To accurately rank a structure, its fitness is determined by the best solution found by particle swarm optimization and numerical integration. The authors of [22] propose a statistical search algorithm called Reactmine to infer chemical reactions with ODE semantics. In [15], the sparse identification of non-linear dynamics (SINDy) [4] is adapted to the stochastic semantics (cf. Section 2). This is achieved by relying on the moment-equations of the chemical master equation, an ODE system describing the time-evolution of the Markov chain's moments. A two-step regression approach, called Reactionet LASSO, is employed to achieve robustness against heteroscedastic, noisy measurements and reaction constants of different magnitudes.

A recent publication adjusts the SINDy approach to accommodate coupled differential equations such as those resulting from the ODE semantics of reaction networks [5]; [12] also

brings SINDy to the case of biochemical systems with mass-action kinetics and account for uncertainty.

In contrast to the above, here we aim at a simulation-based optimization approach, which also allows, e.g., the straightforward inclusion of unmeasured species, arbitrary kinetics, and accounting for probability distributions (instead of their moments). Further, our proposed methods do not rely on numerical differentiation of the time-series data, which can be inaccurate in the presence of noise and large or uneven sampling intervals.

The use of gradient descent for parameter estimation of simulation models also saw great interest recently [1, 6], including biochemical reaction systems [32]. In [33], gradient descent is used to enable Bayesian inference over general ODE models.

5 Learning Reaction Systems with Gradient Descent

Consider a reaction system $\mathbf{R}$ with coefficients $\mathbf{C}$, stochastic rate constants $\mathbf{r} \in \mathbb{R}^n$ and initial populations $\mathbf{S}_{init}$. Assuming the structure $\mathbf{C}$ of the model is known, we can simulate trajectories over states $\mathbf{S}_t, t \geq 0$ from the CTMC, defined by $\mathbf{C}$ and a certain parametrization $\mathbf{r}$. Typically, we want the trajectories produced by $\mathbf{R}$ to resemble the behavior of a reference system. To achieve this, suitable parameter values have to be estimated from collected time-series data: Given measurements D_t at discrete times $t \in \{1, \ldots, n\}$, the goal is to maximize the likelihood $\mathcal{L}(D|\mathbf{r})$ or some other measure of goodness of fit. Determining the parameters $\mathbf{r}$ that maximize the likelihood is also referred to as the inverse problem, since a "forward" simulation provides a sample from $\mathcal{L}$ for a given $\mathbf{r}$.

Here, our goal is to simultaneously infer the structure of the model, i.e., we try to find $\mathbf{r}$ *and* $\mathbf{C}$, such that $\mathcal{L}(D|\mathbf{C}, \mathbf{r})$ is maximal. Obviously, this is a much harder task than just estimating parameters, as the degrees of freedom in the inverse problem are drastically increased. Further, even when taking the mean over SSA trajectories, the response surface may now exhibit jumps, introduced by the discrete entries in $\mathbf{C}$. In fact, we can formulate the problem with varying degrees of smoothness (prior to considering a smoothed objective, cf. Section 3). The following formulations are adapted to the SIR model (cf. Section 2), which we later use for evaluation.

Library of Reactions. Our first problem formulation is inspired by the use of reaction libraries in [5, 15]. This approach can directly be translated to a simulation-based optimization problem: the reaction system to optimize comprises (a selection of) all reactions for a given number of species. The task is to adjust $\mathbf{r}$, where reactions i with r_i below a certain threshold are dropped from the final model. Our library consists of the 36 binary reactions that abide by the conservation law $S + I + R = 2000$. This problem is completely smooth in all dimensions.

Coefficient Steps. In the second problem formulation, we fix the number of reactions to two and try to adjust $\mathbf{C}$ with $c_{ij} \in \{0, 1, 2\}$ and $\mathbf{r}$ directly, yielding a 14-dimensional problem. This problem is non-smooth in the coefficient dimensions.

Reaction Steps. In the third formulation, we again work with a library of reactions but introduce a (continuous) ranking vector of the same dimensionality as $\mathbf{r}$. In each simulation run, only the two reactions with the highest rank are considered, enforcing a certain model size. The task is then to adjust the ranking together with the two rate parameters, one for each reaction in the top-two.

Library of Systems. The final formulation, which we adopt for didactic purposes, is a brute-force approach. It optimizes the 1260 rates of all possible combinations of two reactions from our library of 36 simultaneously. With one optimization run per model being much more performant, this example is designed to showcase the gradient estimator's ability to steer the rate adjustment across large numbers of structures.

Generally, more than one reaction system can produce trajectories from the distribution in D [7]. It is often hard to choose the "right" system automatically, and the choice must be left to domain experts [12]. However, certain criteria can constrain the optimization process to desirable solutions, such as parsimony (choosing a low number of reactions producing a good fit) and background/prior knowledge (such as number of species, conservation laws, or even known reactions). Some of these constraints may result in an NP-hard problem for which the best-known solution is brute force [10]. This can be overcome, e.g., by regularization (like in SINDy) and relaxation.

As we will demonstrate on the example of the problems above, there is a tradeoff between the ability to strongly enforce these constraints and the smoothness of the objective function, which in turn determines the difficulty of the optimization task.

5.1 Reparametrization

In both parameter estimation and structure identification, the scale of the parameters poses a problem: depending on the model, the rate constants can cover multiple orders of magnitude. This is detrimental for many optimization algorithms, as an appropriate step size depends on the dimension of $\mathbf{r}$. This has been tackled in [15] by a separate optimization run to determine the orders of magnitude. The authors of [26] use hand-crafted and learned dilation functions. Here, we use a simple logarithmic reparametrization, which decreases the dynamic range of the parameters:

$$\mathbf{r}' = \exp(a\mathbf{r} + c) - \exp(c), \text{ with } a = \frac{1}{4} \text{ and } c = -20$$

Optimizing in this space means that a step in $\mathbf{r}$ between, e.g., 0.1 and 0.2 is the same as between 1 and 2. The specific shifting and scaling ensure (1) that the value $r_i = 0$ is mapped to 0 and (2) that the values in a sensible range (around 10^{-4} to 10^2) are sufficiently spread. This way, the sensitivity of the response wrt. changes in $\mathbf{r}$ is decreased, aiding the optimization

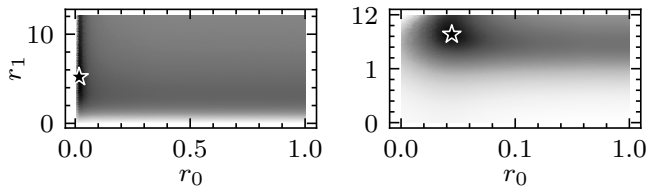

Figure 1: The SIR model's response surface (left) and the effect of reparametrization (right). A darker color equals a lower loss and the star marks the optimum.

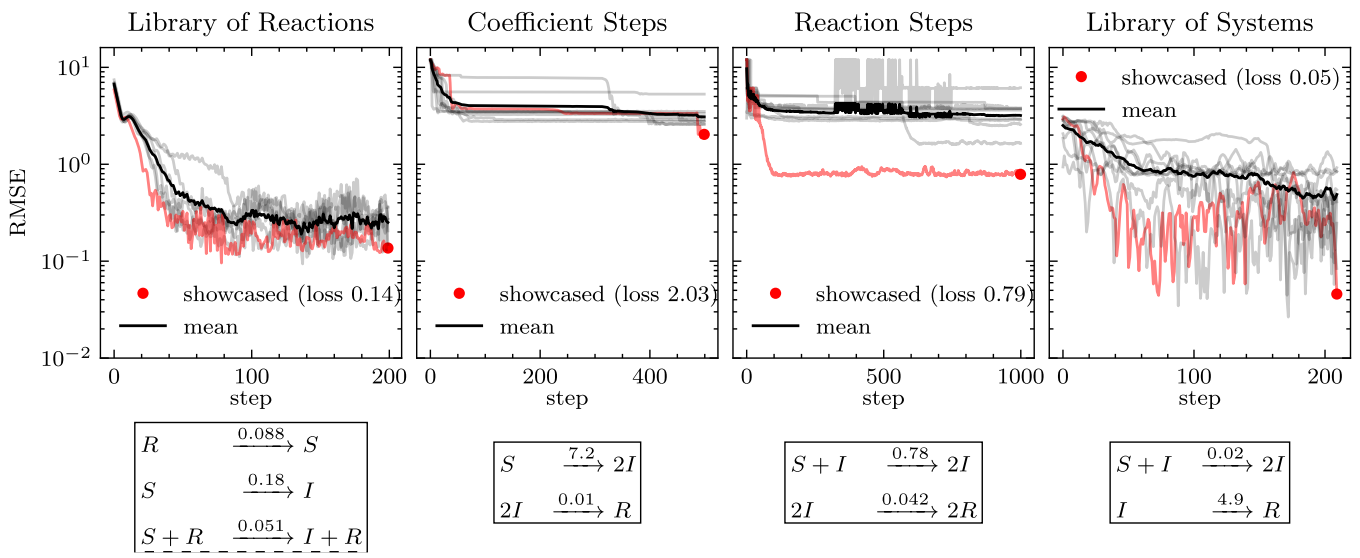

Figure 2: Convergence of gradient descent on the four problems (top) and chosen inferred models (bottom). Progress on the unsmoothed objective, the optimal solution has a loss of about 0.01 (depending on the inferred system's stochasticity). The reaction system shown for *Library of Reactions* comprises 17 reactions above the threshold 10^{-4}, but only the top 3 are shown here.

(cf. Figure 1). Specifically, in the case of our stochastic gradient estimator, this allows us to set a single smoothing factor σ for all dimensions, which would otherwise lead to an overly smoothed objective and occlude narrow minima.

5.2 Evaluation Setup

To identify the challenges and opportunities of gradient descent in the context of a stochastic simulation-based model inference, we evaluate the convergence of our four problem formulations on recovering the SIR model as parametrized in Section 2. Our time-series reference data is generated by simulating the model until $t = 1$ and collecting state snapshots at 100 discrete, equidistant simulation times (although we generally require neither equidistance nor completeness). For optimization, we employ the stochastic gradient estimator introduced in Section 3 and combine it with the Adam gradient descent optimizer [14]. For each problem, we manually determined hyperparameters (sample size n, smoothing factor σ, and learning rate η) that achieved good results. In the order of the problems from Section 5, these are $(100, 0.2, 1)$, $(1000, 1, 1)$, $(100, 0.2, 0.1)$, and $(100, 0.2, 0.5)$. Initial parameters are drawn from problem-specific uniform distributions. Our simple demonstration aims to minimize the root mean squared error (RMSE) between the reference and the simulation time-series, which is run for 20 replications. Note that it is easily possible to change this objective, e.g., to minimizing Wasserstein distances on distribution estimates [28]. We repeat the optimization process 10 times to account for the stochasticity.

6 Results and Discussion

The evaluation results are provided in Figure 2, which shows the mean convergence behavior over gradient descent steps on each problem, as well as the final model inferred by a chosen optimization run. For the Brute Force problem, the lowest RMSE of all structures is shown.

The *Library of Reactions* formulation yields a very precise fit to the input data but lacks parsimony. Convergence is attained fast, as the objective is smooth. Here, a parsimony-encouraging initialization, as with the horseshoe prior for Bayesian regression may prove beneficial [12], albeit introducing bias towards certain solutions.

On *Coefficient Steps*, on the other hand, the smoothed gradient descent struggles to converge to a good solution. Our further experiments showed that convergence to very good solutions is possible, but strongly depends on the initialization. This hints at the existence of hard-to-escape local minima.

In *Reaction Steps*, the smoothed gradient should be able to capture the effects of possible alternate rankings, and we observe good initial progress toward a parsimonious solution. Still, the decoupling of rates and structure seems to be challenging to overcome. When the ranking vector tends to a local minimum, means of escaping it by (partially) shuffling the current ranking could help to identify better solutions in other parts of the search space. However, in preliminary experiments of this sort, we observed inferior results.

Being completely smooth, the brute force *Library of Systems* approach is similar in convergence to the Library of Reactions. In contrast to the latter, it is able to recover the parsimonious original model. This indicates the ability of gradient descent to optimize a vast number of reaction systems at a time. Since the combinatorial explosion puts larger reaction systems out of reach, the main missing building block for this approach is a goal-driven exploration of structures.

Our initial results demonstrate a tradeoff between parsimony, goodness of fit and scalability. This is the result of different response surfaces and their amenability to gradient descent.

In all cases, the scaling of rate constants poses a problem, which can be dealt with by reparametrization (cf. Section 5.1). Whereas the rate constant space clearly places solutions of similar quality close to each other (cf. Figure 1), it is generally

unclear which steps in the structure dimension (e.g., on the coefficients in $\mathbf{C}$) lead to lower loss. The simultaneous adjustment of both $\mathbf{C}$ and $\mathbf{r}$ further complicates solutions that try to (smoothly) enforce a certain model size. A major step towards better convergence would thus be a combined reparametrization of $\mathbf{C}$ and $\mathbf{r}$, which enables a goal-driven exploration of structures. Clearly, such a reparametrization must be approximate, and its existence is unclear, demanding further investigation. Promisingly, in the related case of learning (imperative) programs, first steps have been taken in this direction [19]. Besides parsimony, identifyability could be facilitated by constraining solutions on background knowledge, as for example derived from a conceptual model in a simulation study.

Beyond considering the challenges outlined above, future work may explore the application of other smooth gradient estimation schemes based on automatic differentiation, such as StochasticAD [2] or DiscoGrad [17]. Finally, the full potential of the simulation-based approach needs to be explored, e.g., by considering unmeasured variables and alternative loss functions.

Acknowledgements

JNK and AU acknowledge the funding of the Deutsche Forschungsgemeinschaft under Grant No.: 320435134 (https://gepris.dfg.de/gepris/projekt/320435134); PA is supported by Grant No.: 497901036 (https://gepris.dfg.de/gepris/projekt/497901036).

References

[1] Philipp Andelfinger. "Towards Differentiable Agent-Based Simulation". In: *ACM Trans. Model. Comput. Simul.* 32.4 (Jan. 2023). ISSN: 1049-3301. DOI: 10.1145/3565810.

[2] Gaurav Arya et al. "Automatic Differentiation of Programs with Discrete Randomness". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 10435–10447. URL: https://proceedings.neurips.cc/paper%5C_files/paper/2022/file/43d8e5fc816c692f342493331d5e98fc-Paper-Conference.pdf.

[3] Ehsan Askari and Guillaume Crevecoeur. "Evolutionary sparse data-driven discovery of multibody system dynamics". en. In: *Multibody System Dynamics* 58 (2 June 2023), pp. 197–226. DOI: 10.1007/s11044-023-09901-z.

[4] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". en. In: *Proceedings of the National Academy of Sciences* 113 (15 Apr. 2016), pp. 3932–3937. DOI: 10.1073/pnas.1517384113.

[5] Pamela M. Burrage, Hasitha N. Weerasinghe, and Kevin Burrage. "Using a library of chemical reactions to fit systems of ordinary differential equations to agent-based models: a machine learning approach". en. In: *Numerical Algorithms* (Jan. 2024). DOI: 10.1007/s11075-023-01737-0.

[6] Ayush Chopra et al. *Differentiable Agent-based Epidemiology*. 2023. arXiv: 2207.09714 [cs.LG].

[7] Gheorghe Craciun and Casian Pantea. "Identifiability of chemical reaction networks". In: *Journal of Mathematical Chemistry* 44.1 (2008), pp. 244–259. DOI: 10.1007/s10910-007-9307-x.

[8] Bryan C. Daniels and Ilya Nemenman. "Automated adaptive inference of phenomenological dynamical models". en. In: *Nature Communications* 6 (1 Aug. 2015). DOI: 10.1038/ncomms9133.

[9] Daniel T Gillespie. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". en. In: *Journal of Computational Physics* 22 (4 Dec. 1976), pp. 403–434. DOI: 10.1016/0021-9991(76)90041-3.

[10] Aparna Gupte and Vinod Vaikuntanathan. *The Fine-Grained Hardness of Sparse Linear Regression*. 2022. arXiv: 2106.03131 [cs.LG].

[11] Sayuri K Hahl and Andreas Kremling. "A comparison of deterministic and stochastic modeling approaches for biochemical reaction systems: on fixed points, means, and modes". In: *Frontiers in genetics* 7 (2016), p. 157. DOI: 10.3389/fgene.2016.00157.

[12] Richard Jiang et al. "Identification of dynamic mass-action biochemical reaction networks using sparse Bayesian methods". In: *PLOS Computational Biology* 18.1 (Jan. 2022), pp. 1–21. DOI: 10.1371/journal.pcbi.1009830.

[13] Sarah M Keating et al. "SBML Level 3: an extensible format for the exchange and reuse of biological models". In: *Molecular systems biology* 16.8 (2020), e9110. DOI: https://doi.org/10.15252/msb.20199110.

[14] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: (Dec. 2014). arXiv: 1412.6980v9 [cs.LG].

[15] Anna Klimovskaia, Stefan Ganscha, and Manfred Claassen. "Sparse Regression Based Structure Learning of Stochastic Reaction Networks from Single Cell Snapshot Time Series". en. In: *PLOS Computational Biology* 12 (12 Dec. 2016), e1005234. DOI: 10.1371/journal.pcbi.1005234.

[16] John R. Koza et al. "Reverse Engineering of Metabolic Pathways From Observed Data Using Genetic Programming". In: *Biocomputing 2001*, pp. 434–445. DOI: 10.1142/9789814447362_0043.

[17] Justin N. Kreikemeyer and Philipp Andelfinger. "Smoothing Methods for Automatic Differentiation Across Conditional Branches". In: *IEEE Access* 11 (2023), pp. 143190–143211. DOI: 10.1109/access.2023.3342136.

[18] Thomas G Kurtz. "The relationship between stochastic and deterministic models for chemical reactions". In: *The Journal of Chemical Physics* 57.7 (1972), pp. 2976–2978. DOI: 10.1063/1.1678692.

[19] Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. "Grammar Variational Autoencoder". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 1945–1954. URL: https://proceedings.mlr.press/v70/kusner17a.html.

[20] Wolfgang Maass et al. "Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research". In: *Journal of the Association for Information Systems* 19.12 (2018), p. 1. DOI: 10.17705/1jais.00526.

[21] Charles C. Margossian. "A review of automatic differentiation and its efficient implementation". en. In: *WIREs Data Mining and Knowledge Discovery* 9 (4 July 2019). DOI: 10.1002/widm.1305.

[22] Julien Martinelli et al. *Reactmine: a statistical search algorithm for inferring chemical reactions from time series data*. 2023. arXiv: 2209.03185 [q-bio.QM].

[23] Harley H McAdams and Adam Arkin. "It's a noisy business! Genetic regulation at the nanomolar scale". en. In: *Trends in Genetics* 15 (2 Feb. 1999), pp. 65–69. DOI: 10.1016/s0168-9525(98)01659-x.

[24] Yurii Nesterov and Vladimir Spokoiny. "Random Gradient-Free Minimization of Convex Functions". en. In: *Foundations of Computational Mathematics* 17 (2 Apr. 2017), pp. 527–566. DOI: 10.1007/s10208-015-9296-2.

[25] Marco S. Nobile et al. "Reverse engineering of kinetic reaction networks by means of Cartesian Genetic Programming and Particle Swarm Optimization". In: *2013 IEEE Congress on Evolutionary Computation (CEC)* (Cancun, Mexico). IEEE, June 2013. DOI: 10.1109/cec.2013.6557752.

[26] Marco S. Nobile et al. "Shaping and Dilating the Fitness Landscape for Parameter Estimation in Stochastic Biochemical Models". en. In: *Applied Sciences* 12 (13 July 2022), p. 6671. DOI: 10.3390/app12136671.

[27] Frank Noé, Gianni De Fabritiis, and Cecilia Clementi. "Machine learning for protein folding and dynamics". In: *Current opinion in structural biology* 60 (2020), pp. 77–84. DOI: 10.1016/j.sbi.2019.12.005.

[28] Kaan Öcal, Ramon Grima, and Guido Sanguinetti. "Parameter estimation for biochemical reaction networks using Wasserstein distances". In: *Journal of Physics A: Mathematical and Theoretical* 53 (3 Jan. 2020), p. 034002. DOI: 10.1088/1751-8121/ab5877.

[29] B.T. Polyak. *Introduction to Optimization*. New York: Optimization Software, 1987.

[30] Rajesh Ramaswamy et al. "Discreteness-induced concentration inversion in mesoscopic chemical systems". en. In: *Nature Communications* 3 (1 Apr. 2012). DOI: 10.1038/ncomms1775.

[31] Wen Jun Tan, Moon Gi Seok, and Wentong Cai. "Automatic Model Generation and Data Assimilation Framework for Cyber-Physical Production Systems". In: *Proceedings of the 2023 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. SIGSIM-PADS '23. Orlando, FL, USA: Association for Computing Machinery, 2023, pp. 73–83. ISBN: 9798400700309. DOI: 10.1145/3573900.3591112.

[32] Yuanfeng Wang et al. "Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent". In: *BMC systems biology* 4.1 (2010), pp. 1–16. DOI: 10.1186/1752-0509-4-99.

[33] Yibo Yang, Mohamed Aziz Bhouri, and Paris Perdikaris. "Bayesian differential programming for robust systems identification under uncertainty". In: *Proceedings of the Royal Society A* 476.2243 (2020), p. 20200290. DOI: https://doi.org/10.1098/rspa.2020.0290.