Deep Grey-box Modeling With Adaptive Data-Driven Models Toward Trustworthy Estimation of Theory-Driven Models

Naoya Takeishi and **Alexandros Kalousis**

Geneva School of Business Management, University of Applied Sciences and Arts Western Switzerland (HES-SO)

Take-Home Messages

- Deep grey-box models = theory + DNN
- Not surprisingly, learning deep grey-box models needs **regularization**
- But... we don't know the property of regularizer = the property of the estimation of theory model parameters, θ_T
- We should **empirically analyze the behavior of regularizers**; to this end, marginalizing out θ_T helps

Grey-box Modeling

Combination of data-driven models (eg NNs) & theory-driven models (first principles or expert's experiences) may be advantageous in terms of:

- sample complexity;
- extrapolation performance;
- interpretability.

(Though, we don't really know... it is to be studied!)

We refer to such models as **grey-box models**.

Example For regression from x to y, a simple (yet useful) grey-box model is an additive model:

$$y = f_T(x;\theta_T) + f_D(x;\theta_D),$$

f_T: theory-driven model parameterized by θ_T,
f_D: data-driven model parametrized by θ_D.

We can represent *general* grey-box models as

$$y = C(f_T, f_D; x),$$

where C is a functional that evaluates f_T and f_D on x and mixes their outputs in *some* way.

Learning **Deep** Grey-box Models Needs Regularization

Suppose that f_D is a universal approximator (eg DNN) and that the overall model C attains that property. Then, **empirical risk minimization (ERM) cannot solely choose f_T's parameter, θ_T.**

Example Suppose $C(f_T, f_D; x) = f_T(x;\theta_T) + f_D(x;\theta_D)$ with loss $L(\theta_T,\theta_D) = \sum \|y - C(f_T, f_D; x)\|_2^2$.
If f_D is DNN, it can fit $y - f_T(x;\theta_T)$ for any θ_T; the loss on a training set can be small to the same extent.

Which regularizers should we use? → depends on user's belief; to know the best one *a priori* is difficult.

E.g., $R_{\text{normD}}(\theta_T,\theta_D) := \|f_D(x;\theta_D)\|_{\mathcal{F}_D}^2$ "f_D should act minimally in terms of norm"
$R_{\text{corr}}(\theta_T,\theta_D) := |\langle f_T, f_D\rangle|$ "Values of f_T and f_D should be uncorrelated" ... etc.

Estimated value of θ_T matters (contrarily, value of θ_D does not really matter).
= Property of R matters. Existence of clear minima? How many local minima? etc. (we don't know...)

Proposed Method: Deep Grey-box Models with **Adaptive** Data-Driven Models

Minimizing $L(\theta_T,\theta_D) + \lambda R(\theta_T,\theta_D)$ may be troublesome. Instead, we suggest the following procedures:

1. Make f_D adaptive to the values of θ_T and $f_T(x;\theta_T)$; i.e., $f_D(x;\theta_T) \to f_D(x,\theta_T,f_T(x;\theta_T);\theta_D)$
2. Minimize $\mathbb{E}_{p(\theta_T)}[L(\theta_T,\theta_D)]$ wrt. θ_D only; i.e., marginalize out θ_T
3. You can now evaluate the value of *any* R for *any* θ_T *w/o re-training* because f_D works adaptively to θ_T
4. (optional) Minimize $R(\theta_T,\theta_D^*)$ wrt. θ_T

Experiments

Data are simulated from the 2D reaction-diffusion system:

$$\partial u/\partial t = 0.0015\Delta u + u - u^3 - v + 0.005,$$
$$\partial v/\partial t = 0.005\Delta v + u - v.$$

Task is to predict u, v for $t \in [1, 15]$ given u, v at $t = 0$.
Model is $f_T + f_D$; $f_T = [a\Delta u, b\Delta u]$ (a, b unknown), and f_D is a CNN (whose filters parameterized also by θ_T).

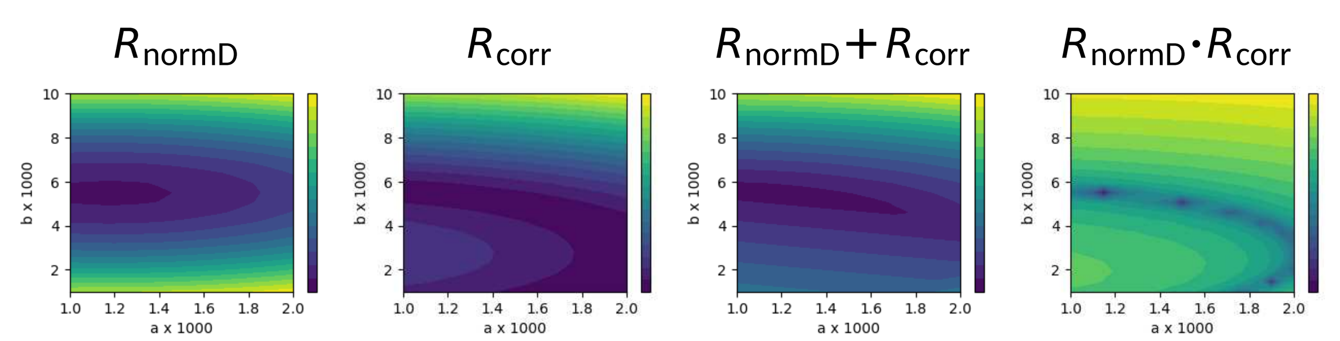

Figure: Landscapes of regularizers. Axes correspond to a and b of f_T.
By the way, the test RMSE is similarly small for any values of a, b.

Data are time-series population densities of prey (algae) and predator (rotifer).
Task is to auto-encode the subsequences.
Model is $f_T + f_D$; f_T is the Lotka–Volterra ($\alpha,\beta,\gamma,\delta$ unknown), and f_D is an MLP.

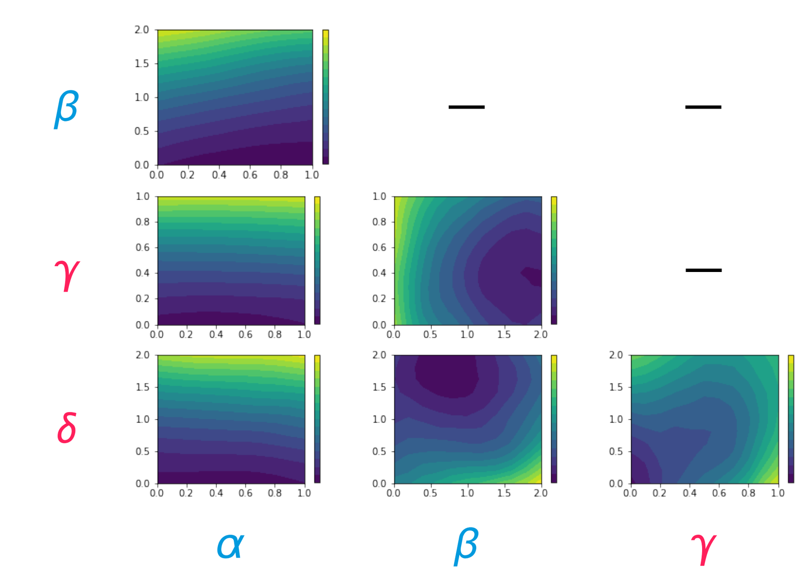

Figure: Landscape of R_{normD} at some timestep.

Discussions

- We don't suggest any regularizers; it needs to be discussed by practitioners
- Our suggestion provides a way for exploratory data analysis. Care must be taken so that data are not reused inappropriately
- A byproduct of the proposed formulation is the decoupled optimization of θ_D and θ_T

References

[1] Y. Yin, V. Le Guen, J. Dona, E. de Bézenac, I. Ayed, N. Thome, and P. Gallinari.
Augmenting physical models with deep networks for complex dynamics forecasting.
In *ICLR*, 2021.

[2] Z. Qian, W. R. Zame, L. M. Fleuren, P. Elbers, and M. van der Schaar.
Integrating expert ODEs into neural ODEs: Pharmacology and disease progression.
In *NeurIPS*, pages 11364–11383, 2021.

[3] N. Takeishi and A. Kalousis.
Physics-integrated variational autoencoders for robust and interpretable generative modeling.
In *NeurIPS*, pages 14809–14821, 2021.

[4] A. Wehenkel, J. Behrmann, H. Hsu, G. Sapiro, G. Louppe, and J.-H. Jacobsen.
Robust hybrid learning with expert augmentation.
TMLR, 2023.

Acknowledgements

This work was supported by Innosuisse (39453.1 IP-ICT) and SNSF Sinergia (CRSII5_177179) projects.

Contact

https://ntake.jp/ and http://dmml.ch/
Updated poster/paper, if any, is available there 😸