NEURAL INFORMATION PROCESSING SYSTEMS

Incorporating Surrogate Gradient Norm to Improve Offline Optimization Techniques

Manh Cuong Dao[1], Phi Le Nguyen[1], Thao Nguyen Truong[2],
Trong Nghia Hoang[3]

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY [1]

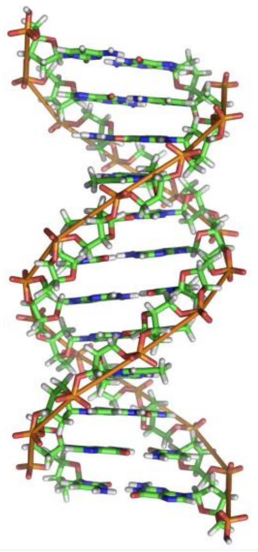
AIST [2]

WASHINGTON STATE UNIVERSITY [3]

Problem Definition

- Find a design $\boldsymbol{x}$ that maximizes certain desirable properties.
  - For instance:
    - Find a DNA sequence with maximum binding affinity.

[1]



- However, evaluation $g(\boldsymbol{x})$ is prohibitively expensive.
  - For instance:
    - Expensive laboratory experiment to measure binding affinity.

- **Offline Model-based Optimization (MBO):** Given an offline dataset $\mathfrak{D} = (\boldsymbol{x}_i, z_i)_{i=1}^{n}$ where $z_i = g(\boldsymbol{x}_i)$ with $g(.)$ is an **unknown** oracle function, find

$$\boldsymbol{x}_* = \underset{\boldsymbol{x} \in \mathcal{X}}{\operatorname{argmax}}\, g(\boldsymbol{x})$$

[1] Brandon Trabucco, Xinyang Geng, Aviral Kumar, and Sergey Levine. Design-bench: Benchmarks for data-driven offline model-based optimization. In International Conference on Machine Learning, pages 21658–21676. PMLR, 2022.

A direct approach to MBO

- Learn a surrogate $g(\boldsymbol{x};\boldsymbol{\omega}_*)$ of $g(\boldsymbol{x})$ via fitting to the offline dataset.

$$\boldsymbol{\omega}_* = \underset{\boldsymbol{\omega}}{\operatorname{argmin}}\, \mathcal{L}_{\mathfrak{D}}(\boldsymbol{\omega})$$

- The (oracle) maxima of $g(\boldsymbol{x})$ is then approximated via:

$$\boldsymbol{x}_* = \underset{\boldsymbol{x}}{\operatorname{argm}ax}\, g(\boldsymbol{x};\boldsymbol{\omega}_*)$$

- **Challenge:** Predictions of $g(\boldsymbol{x};\boldsymbol{\omega}_*)$ are **unreliable** in OOD regime.

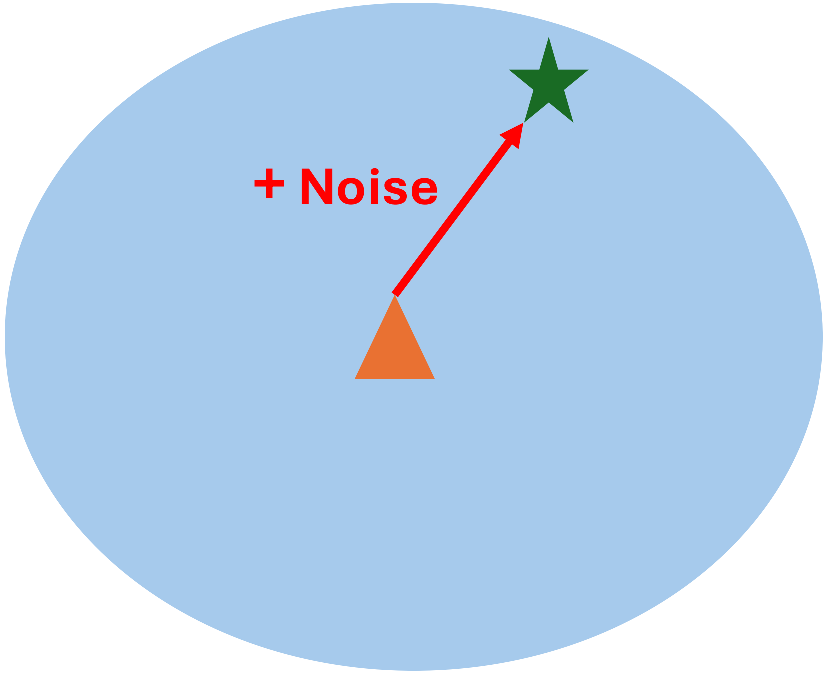
[2] Brandon Trabucco, Aviral Kumar, Xinyang Geng, and Sergey Levine. Conservative objective models for effective offline model-based optimization. In International Conference on Machine Learning, pages 10358–10368. PMLR, 2021.

Motivation

- Suppose:

Oracle $g(.)$ lies within the parametric family of Surrogate $g(.;\boldsymbol{\omega}_*)$

there must exist a perturbation neighborhood of Surrogate $g(.;\boldsymbol{\omega}_*)$ that contains Oracle $g(.)$



Oracle $g(.)$

Surrogate $g(.;\boldsymbol{\omega}_*)$

Perturbation neighborhood of surrogate

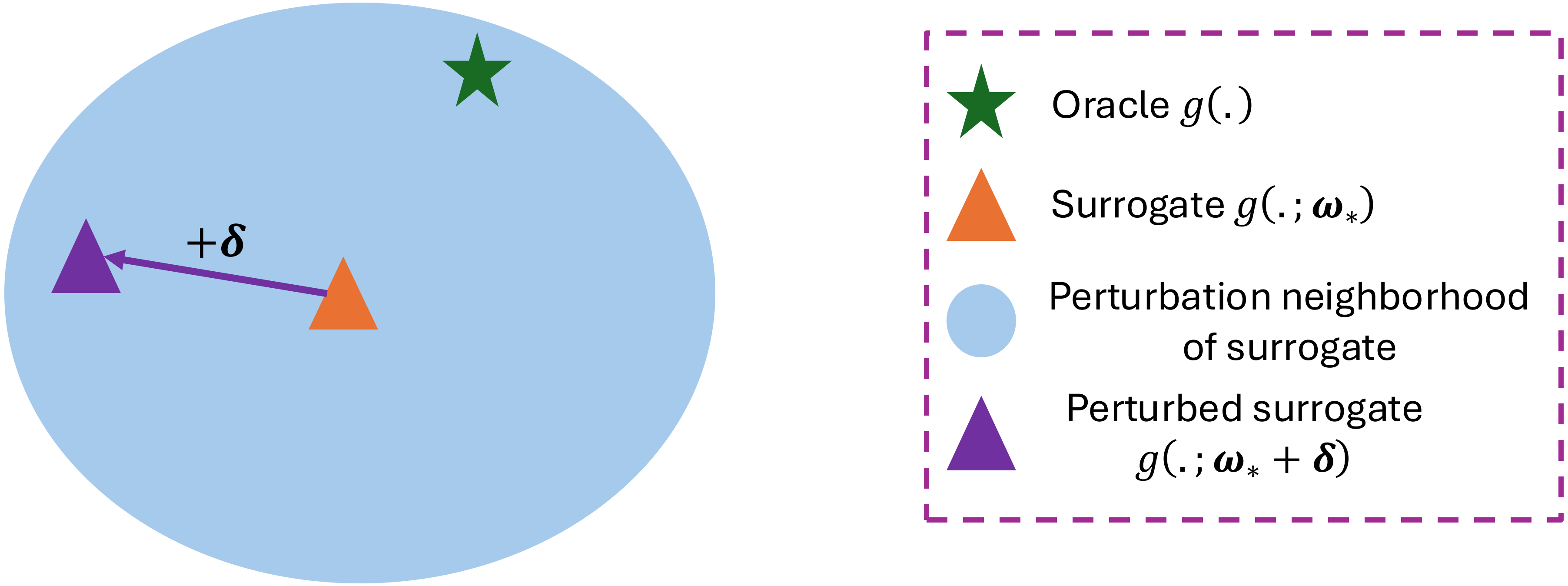
Noise vector to obtain oracle from surrogate

Motivation

- Suppose:

Prediction of $g(.;\boldsymbol{\omega}_* + \boldsymbol{\delta})$ do not change substantially on

Surrogate's prediction $\approx$ Oracle's prediction



Find surrogate s.t. worst-case prediction change across the perturbation neighborhood is sufficiently small.

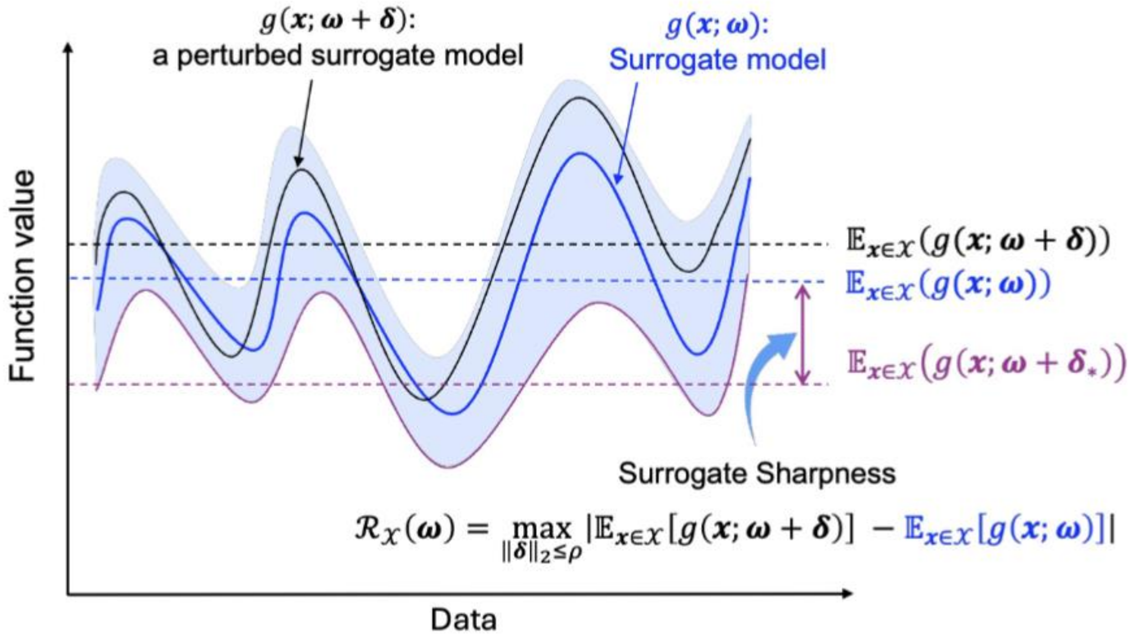
Surrogate sharpness

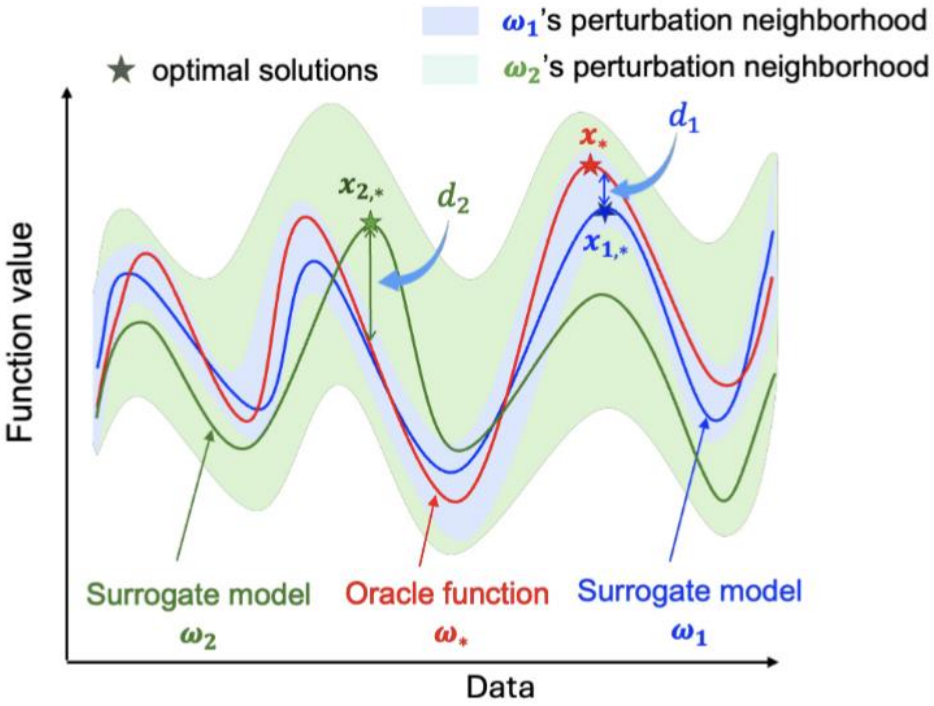
- **Surrogate sharpness:**

$$\color{red}{\mathcal{R}_{\mathcal{X}}(\boldsymbol{\omega})} = \max_{\|\boldsymbol{\delta}\|_2 < \rho} \left| \mathbb{E}_{x \in \mathcal{X}}[g(x; \boldsymbol{\omega} + \boldsymbol{\delta})] - \mathbb{E}_{x \in \mathcal{X}}[g(x; \boldsymbol{\omega})] \right|$$

- This can be used to regularize surrogate training:

$$\boldsymbol{\omega}_* = \operatorname*{argmin}_{\boldsymbol{\omega}} \mathcal{L}_{\mathfrak{D}}(\boldsymbol{\omega}) \quad \text{s.t.} \quad \color{red}{\mathcal{R}_{\mathcal{X}}(\boldsymbol{\omega})} \le \epsilon'$$





Surrogate sharpness

- We proved in Theorem 2 that

$$\mathcal{R}_{\mathcal{X}}(\boldsymbol{\omega}) \leq \left(\rho G(\boldsymbol{\omega}) + \frac{1}{2}\rho^2 \lambda_{max}\right) \cdot \left(\mathcal{R}_{\mathcal{D}}(\boldsymbol{\omega}) + \mathcal{O}\left(\sqrt{\frac{\dim(\boldsymbol{\omega}) \log(n\|\boldsymbol{\omega}\|^2)}{n}}\right)\right)$$

where $G(\boldsymbol{\omega}) = \|\mathbb{E}_{x\in\mathcal{X}}[\nabla_{\boldsymbol{\omega}} g(x;\boldsymbol{\omega})]\|$ and λ_{max} is largest eigenvalue of Hessian of the surrogate's expected prediction.

This can transform the constraint:

$$\boldsymbol{\omega}_* = \operatorname*{argmin}_{\boldsymbol{\omega}} \mathcal{L}_{\mathfrak{D}}(\boldsymbol{\omega}) \quad \text{s.t.} \quad \mathcal{R}_{\mathcal{D}}(\boldsymbol{\omega}) \leq \epsilon$$

Practical Algorithms

- Let $h(\boldsymbol{\omega}+\boldsymbol{\delta}) = \mathbb{E}_{x\in\mathcal{D}}[g(x;\boldsymbol{\omega}+\boldsymbol{\delta})]$, $h(\boldsymbol{\omega}) = \mathbb{E}_{x\in\mathcal{D}}[g(x;\boldsymbol{\omega})]$
- Use first-order **Taylor expansion** of $h(\boldsymbol{\omega}+\boldsymbol{\delta})$ at $\boldsymbol{\omega}$:

$$\mathcal{R}_{\mathcal{D}}(\boldsymbol{\omega}) = \max_{\|\boldsymbol{\delta}\|_2<\rho} |\mathbb{E}_{x\in\mathcal{D}}[g(x;\boldsymbol{\omega}+\boldsymbol{\delta})] - \mathbb{E}_{x\in\mathcal{D}}[g(x;\boldsymbol{\omega})]|$$
$$= \max_{\|\boldsymbol{\delta}\|_2<\rho} |h(\boldsymbol{\omega}+\boldsymbol{\delta}) - h(\boldsymbol{\omega})| \approx \max_{\|\boldsymbol{\delta}\|_2<\rho} |\nabla_{\boldsymbol{\omega}} h(\boldsymbol{\omega})^T \boldsymbol{\delta}|$$

- Use the **Cauchy-Schwartz** inequality:

$$\mathcal{R}_{\mathcal{D}}(\boldsymbol{\omega}) \approx \max_{\|\boldsymbol{\delta}\|_2<\rho} |\nabla_{\boldsymbol{\omega}} h(\boldsymbol{\omega})^T \boldsymbol{\delta}| = \max_{\|\boldsymbol{\delta}\|_2<\rho} \|\nabla_{\boldsymbol{\omega}} h(\boldsymbol{\omega})\| . \|\boldsymbol{\delta}\| = \rho . \|\nabla_{\boldsymbol{\omega}} h(\boldsymbol{\omega})\|$$

- Surrogate training can be rewritten as:

$$\boldsymbol{\omega}_* = \operatorname*{argmin}_{\boldsymbol{\omega}} \mathcal{L}_{\mathfrak{D}}(\boldsymbol{\omega}) \quad \text{s.t.} \quad \rho . \|\nabla_{\boldsymbol{\omega}} h(\boldsymbol{\omega})\| \le \epsilon$$

- This can be solved via **Lagrangian**:

$$\boldsymbol{\omega}_* = \operatorname*{argmin}_{\boldsymbol{\omega}} \mathcal{L}(\boldsymbol{\omega},\lambda) \quad \text{where} \quad \mathcal{L}(\boldsymbol{\omega},\lambda) = \mathcal{L}_{\mathfrak{D}}(\boldsymbol{\omega}) + \lambda(\rho . \|\nabla_{\boldsymbol{\omega}} h(\boldsymbol{\omega})\| - \epsilon)$$

Practical Algorithms

- Utilize the **basic differential multiplier method** (**BDMM**)[3], which simultaneously**:**
  - Gradient descent for $\boldsymbol{\omega}$:

$$\boldsymbol{\omega}^{t+1} = \boldsymbol{\omega}^{t} - \eta_{\boldsymbol{\omega}}.\left(\nabla_{\boldsymbol{\omega}}\mathcal{L}_{\mathfrak{D}}(\boldsymbol{\omega}) + \lambda^{t}.\rho.\nabla_{\boldsymbol{\omega}}\|\nabla_{\boldsymbol{\omega}}h(\boldsymbol{\omega}^{t})\|\right)$$

  - Gradient ascent for λ:

$$\lambda^{t+1} = \lambda^{t} + \eta_{\lambda}.\left(\rho.\|\nabla_{\boldsymbol{\omega}}h(\boldsymbol{\omega})\| - \epsilon\right)$$

➡ We name this method **IGNITE**.

[3] John Platt and Alan Barr. Constrained differential optimization. In Neural Information Processing Systems, 1987

Experiments

Algorithms		Ant Morphology Performance	Ant Morphology Gain	D'Kitty Morphology Performance	D'Kitty Morphology Gain	TF Bind 8 Performance	TF Bind 8 Gain	TF Bind 10 Performance	TF Bind 10 Gain
$\mathfrak{D}$(best)		0.565		0.884		0.565		0.884	
REINF-ORCE	Base	0.255 ± 0.036		0.546 ± 0.208		0.929 ± 0.043		0.635 ± 0.028	
	IGNITE	0.282 ± 0.021	+2.7%	0.642 ± 0.160	+9.6%	0.944 ± 0.030	+1.5%	0.670 ± 0.060	+3.5%
GA	Base	0.303 ± 0.027		0.881 ± 0.016		0.980 ± 0.016		0.651 ± 0.033	
	IGNITE	0.320 ± 0.044	+1.7%	0.886 ± 0.017	+0.5%	0.985 ± 0.010	+0.5%	0.653 ± 0.043	+0.2%
ENS-MEAN	Base	0.376 ± 0.060		0.888 ± 0.010		0.985 ± 0.009		0.649 ± 0.036	
	IGNITE	0.435 ± 0.058	+5.9%	0.896 ± 0.013	+0.8%	0.987 ± 0.007	+0.2%	0.662 ± 0.091	+1.3%
ENS-MIN	Base	0.385 ± 0.067		0.889 ± 0.014		0.980 ± 0.012		0.681 ± 0.095	
	IGNITE	0.468 ± 0.062	+8.3%	0.897 ± 0.010	+0.8%	0.986 ± 0.010	+0.6%	0.705 ± 0.118	+2.4%
CbAS	Base	0.854 ± 0.042		0.895 ± 0.012		0.919 ± 0.044		0.635 ± 0.041	
	IGNITE	0.859 ± 0.039	+0.5%	0.900 ± 0.015	+0.5%	0.921 ± 0.042	+0.2%	0.652 ± 0.055	+1.7%
MINs	Base	0.905 ± 0.023		0.944 ± 0.008		0.892 ± 0.046		0.643 ± 0.062	
	IGNITE	0.911 ± 0.024	+0.6%	0.945 ± 0.007	+0.1%	0.930 ± 0.041	+3.8%	0.647 ± 0.058	+0.4%
RoMA	Base	0.569 ± 0.086		0.821 ± 0.019		0.665 ± 0.000		0.550 ± 0.008	
	IGNITE	0.615 ± 0.085	+4.6%	0.834 ± 0.012	+1.3%	0.665 ± 0.000	+0.0%	0.553 ± 0.000	+0.3%
COMs	Base	0.897 ± 0.031		0.931 ± 0.013		0.955 ± 0.030		0.645 ± 0.038	
	IGNITE	0.901 ± 0.030	+0.4%	0.934 ± 0.010	+0.3%	0.952 ± 0.043	-0.3%	0.638 ± 0.053	-0.7%
CMA-ES	Base	1.955 ± 1.484		0.724 ± 0.002		0.928 ± 0.040		0.668 ± 0.035	
	IGNITE	1.957 ± 1.910	+0.2%	0.724 ± 0.001	+0.0%	0.927 ± 0.043	-0.1%	0.673 ± 0.044	+0.5%
BO-qEI	Base	0.812 ± 0.000		0.896 ± 0.000		0.787 ± 0.112		0.628 ± 0.000	
	IGNITE	0.812 ± 0.000	+0.0%	0.896 ± 0.000	+0.0%	0.843 ± 0.109	+5.6%	0.628 ± 0.000	+0.0%
ICT	Base	0.937 ± 0.023		0.946 ± 0.014		0.892 ± 0.055		0.647 ± 0.025	
	IGNITE	0.935 ± 0.032	-0.2%	0.962 ± 0.018	+1.6%	0.923 ± 0.038	+3.1%	0.652 ± 0.074	+0.5%

IMPROVE:

- 79.55% = 35/44 cases
- Average improvement: 1.91%
- Peak improvement: 9.6%.

DECREASE:

- 9.09% = 4/44 cases
- Average degradation: 0.3%
- Peak degradation: 0.7%.

MAINTAIN:

- 11.36% = 5/44 cases

Thank you

for listening