On the **Optimality Gap** of Warm-started Hyperparameter Optimization

—

Parikshit Ram
IBM Research AI

IBM

Hyperparameter Optimization

Data domain & distribution $(x, y) \in X \times Y \sim D$

Per-sample loss $\ell : Y \times Y \longrightarrow \mathbb{R}_+$

Model for HP trained on data $f_{\theta,S} : X \longrightarrow Y$

Loss for a HP configuration $L(\theta, D) := \mathbb{E}_{S \sim D^n} \mathbb{E}_{(x,y) \sim D} \ell(y, f_{\theta,S}(x))$

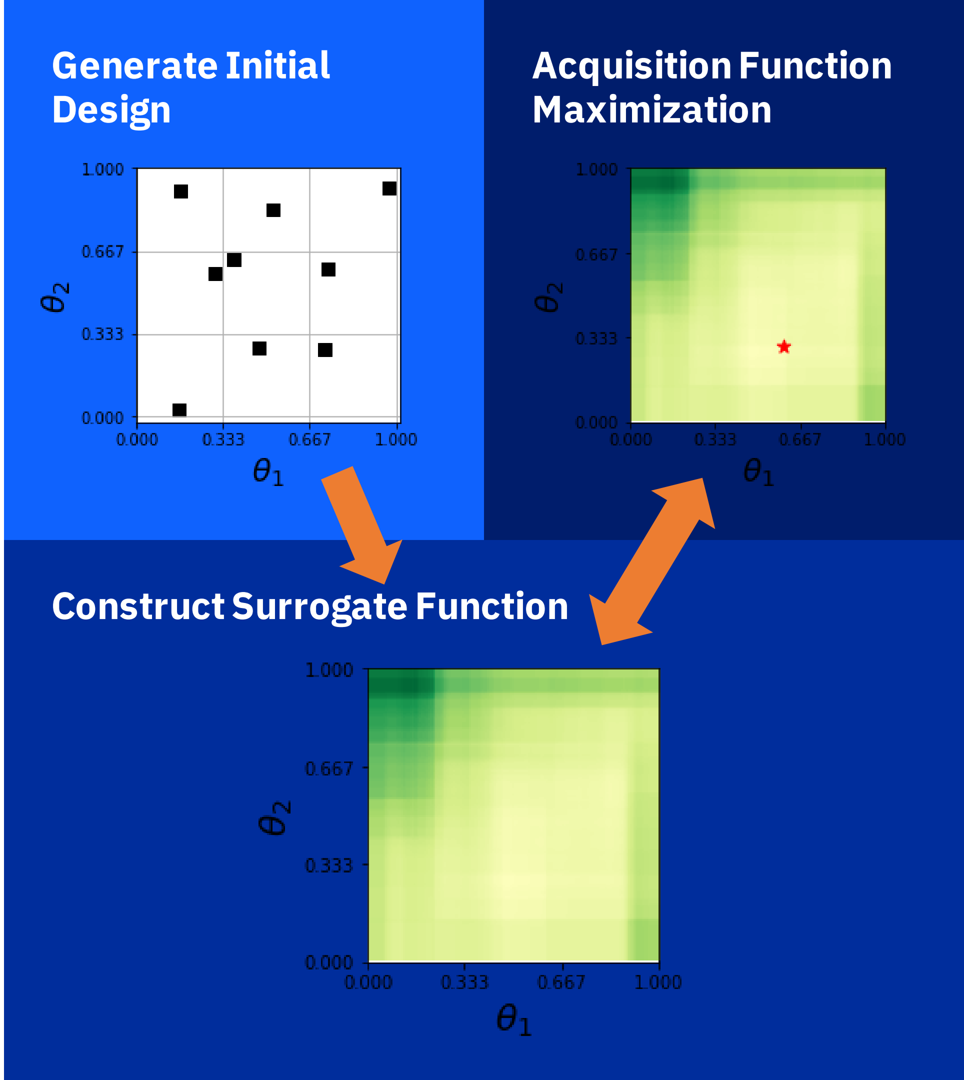
HPO problem

$$\min_{\theta \in \Theta} L(\theta, D)$$

SMBO

Sequential Model Based Optimization

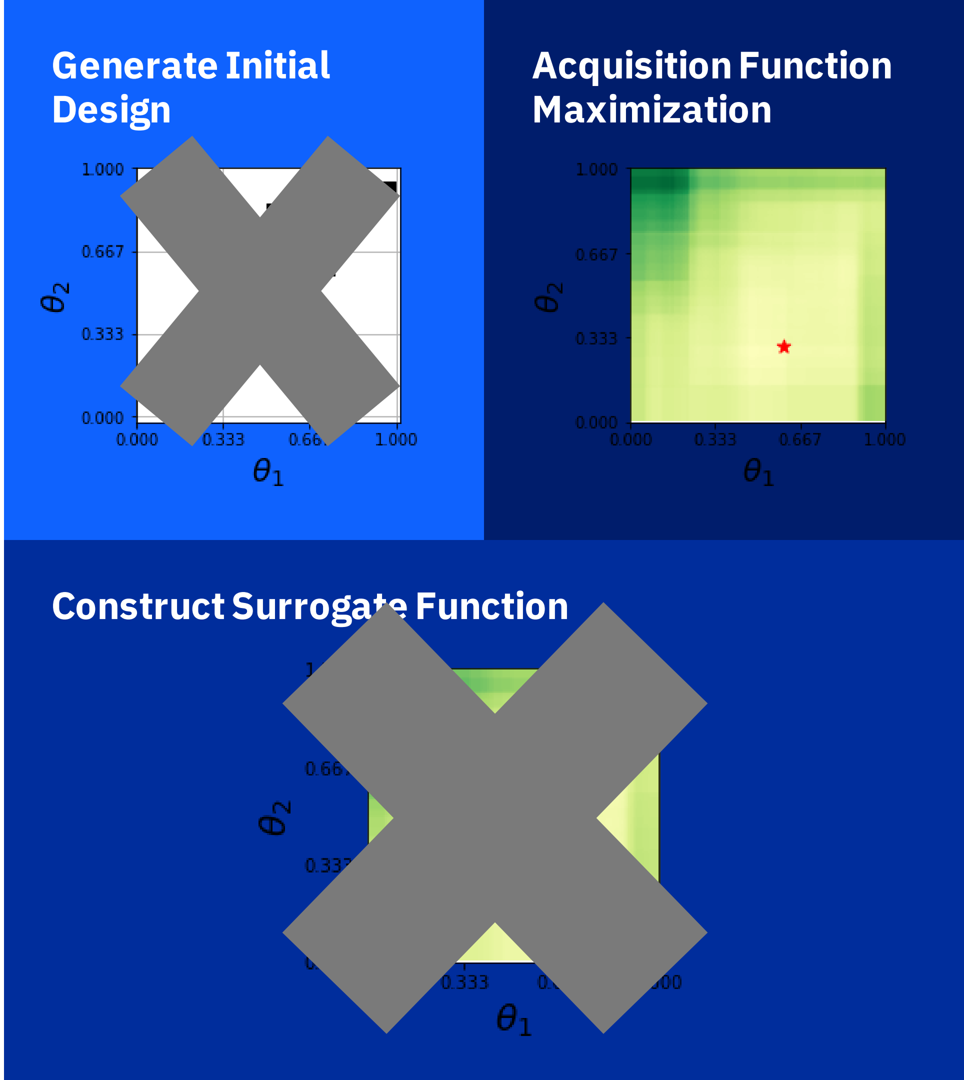
- Generate Initial Design of HPs
- Evaluate HPs
- Until budget expires
  - Construct surrogate model & acquisition function
  - Select next HP via AF maximization
  - Evaluate new HP
  - Add (HP, loss) to set of evaluated HPs
- Select best HP from evaluated set



Few-shot HPO

Very low-budget SMBO

- ~~Generate Initial Design of HPs~~
- Evaluate HPs
- Until budget expires
  - ~~Construct surrogate model~~ & acquisition function
  - Select next HP via AF maximization
  - Evaluate new HP
  - Add (HP, loss) to set of evaluated HPs
- Select best HP from evaluated set



Meta-learning from Previous HPO Experiences

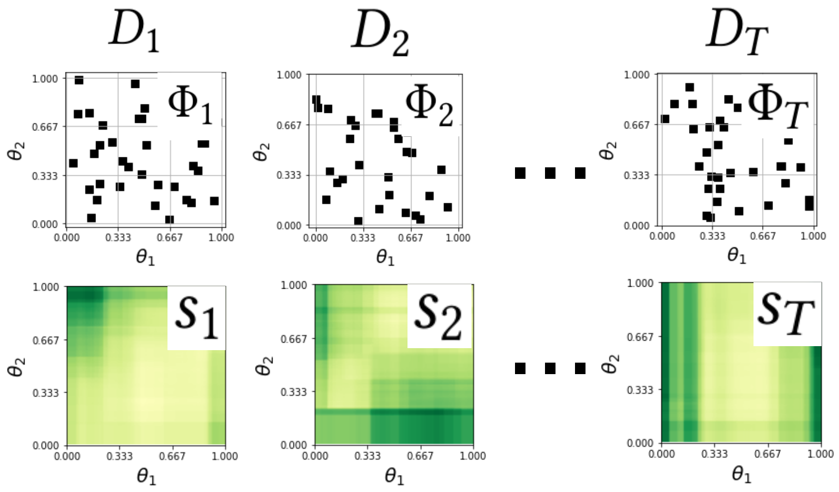
Source tasks:

- Evaluated HPs

$$\Phi_t, t \in [T]$$

- Surrogate functions

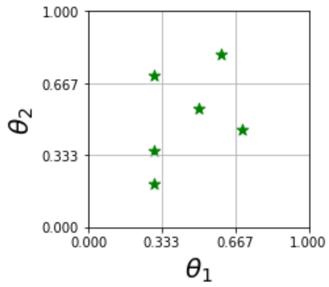
$$s_t, t \in [T]$$



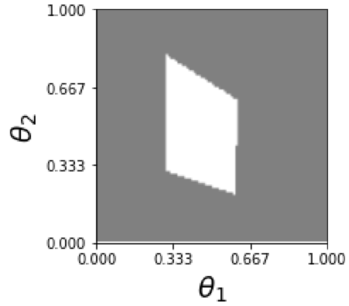
Meta-learning from Previous HPO Experiences

Target task few-shot warm-started HPO

- Meta-learning **small initial design**

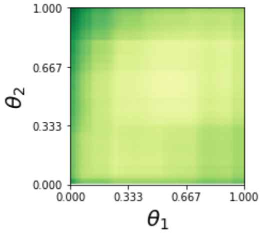


- Meta-learning **pruned HP search space**



- Transfer surrogate functions

$$s(\theta) = \sum_{t \in [T]} \alpha_t(\theta) s_t(\theta)$$



Goal of Analysis & Pre-requisites

Optimality gap upper bound

$$L(\hat{\theta}, D) - L(\theta^\star, D)$$

$\hat{\theta}$ **HP from few-shot HPO**

$\theta^\star$ **Optimal HP for problem**

- Smooth per-sample loss w.r.t. label

  $$\ell : Y \times Y \rightarrow \mathbb{R}_+$$

- Smooth loss w.r.t. HPs

  $$|L(\theta, D) - L(\theta', D)| \leq \gamma \cdot \|\theta - \theta'\|$$

- Smooth surrogate functions w.r.t. HPs

  $$|s_t(\theta) - s_t(\theta')| \leq \omega \cdot \|\theta - \theta'\|$$

- Quantify "domain-gap" between source & target

  $$|L(\theta, D) - L(\theta, D')|$$

Quantifying Domain Gap

Domain gap bound

$$|L(\theta, D) - L(\theta, D')| \leq \beta \cdot W_1 \left(P_\theta(D), P_\theta(D')\right)$$

- Distribution of interest

$$(z_1, z_2) \sim P_\theta(D)$$
$$\Rightarrow (x, y) \sim D, S \sim D^n,$$
$$z_1 \leftarrow y, z_2 \leftarrow f_{\theta,S}(x)$$

- Domain gap is HP specific.
- **No need for distance between different multi-dimensional data distributions of different sizes and dimensionalities; simple 1-Wasserstein distance suffices.**

Best Achievable

$$L(\hat{\theta}, D) - L(\theta^\star, D) \leq \tilde{O}\left(\max_{\theta \in \Theta} \min_{t \in [T]} W_1(P_\theta(D), P_\theta(D_t))\right)$$

Best Achievable

$$L(\hat{\theta}, D) - L(\theta^\star, D) \leq \tilde{O}\left(\max_{\theta \in \Theta} \min_{t \in [T]} W_1(P_\theta(D), P_\theta(D_t))\right)$$

$$\exists t \in [T], D \approx D_t \Rightarrow L(\hat{\theta}, D) \approx L(\theta^\star, D)$$

AutoML 2022 / Baltimore, USA / © 2022 IBM Corporation

Best Achievable

$$L(\hat{\theta}, D) - L(\theta^\star, D) \leq \tilde{O}\left(\max_{\theta \in \Theta} \min_{t \in [T]} W_1(P_\theta(D), P_\theta(D_t))\right)$$

$$\exists t \in [T], D \approx D_t \Rightarrow L(\hat{\theta}, D) \approx L(\theta^\star, D)$$

$$L(\hat{\theta}, D) \approx L(\theta^\star, D) \nRightarrow \exists t \in [T], D \approx D_t$$

Best Achievable

$$L(\hat{\theta}, D) - L(\theta^\star, D) \leq \tilde{O}\left(\max_{\theta \in \Theta} \min_{t \in [T]} W_1(P_\theta(D), P_\theta(D_t))\right)$$

$$\exists \Theta_t \subset \Theta, t \in [T], \cup_{t \in [T]} \Theta_t = \Theta,$$

$$\max_{\theta \in \Theta_t} W_1(P_\theta(D), P_\theta(D_t)) \approx 0$$

$$\Rightarrow L(\hat{\theta}, D) \approx L(\theta^\star, D)$$

Possible to get zero optimality gap without requiring the target distribution to match one of the source distributions

Optimality Gap for Pruned Search Spaces

$$L(\hat{\theta}; D) - L(\theta^\star; D)$$

$$\leq \min_{t \in [T] : \phi_t \in \bar{\Theta}} \left(\gamma \cdot \max_{\theta \in \bar{\Theta}} \|\theta - \phi_t\| + 2\beta \cdot \max_{\theta \in \Theta} W_1 \left(P_\theta(D), P_\theta(D_t) \right) \right)$$

- Smaller pruned spaces help
- Best possible

$$\tilde{O} \left(\min_{t \in [T]} \max_{\theta \in \Theta} W_1 \left(P_\theta(D), P_\theta(D_t) \right) \right)$$

- Zero optimality gap only if target distribution matches one of the source distributions
- Weaker than

$$\tilde{O} \left(\max_{\theta \in \Theta} \min_{t \in [T]} W_1 \left(P_\theta(D), P_\theta(D_t) \right) \right)$$

Optimality Gap for Pruned Search Spaces

$$L(\hat{\theta}; D) - L(\theta^\star; D)$$

$$\leq \min_{t \in [T] : \phi_t \in \bar{\Theta}} \left(\gamma \cdot \max_{\theta \in \bar{\Theta}} \|\theta - \phi_t\| + 2\beta \cdot \max_{\theta \in \Theta} W_1 \left(P_\theta(D), P_\theta(D_t) \right) \right)$$

Meta-learned Initial Design

Meta-learned Bounding Box

Meta-learning Convex Hull

Optimality Gap for Surrogate Transfer

$$s(\theta) := \sum_{t\in[T]} \alpha_t(\theta) \cdot s_t(\theta)$$

– Weights

- Fixed $\alpha_t(\theta) = 1/T \, \forall \theta \in \Theta, \forall t \in [T]$
- Adaptive $\alpha_t(\theta) = \begin{cases} 1, & t = \arg\max_{j\in[T]} s_t(\theta) \\ 0, & \text{o.w.} \end{cases}$

$$L(\hat{\theta}; D) - L(\theta^\star; D)$$
$$\leq 2 \max_{\theta\in\Theta} \sum_{t\in[T]} \alpha_t(\theta) \left(\beta \cdot W_1 \left(P_\theta(D), P_\theta(D_t)\right) + |L(\theta, D_t) - s_t(\theta)|\right)$$

Optimality Gap for Surrogate Transfer

$$L(\hat{\theta}; D) - L(\theta^\star; D)$$
$$\leq 2 \max_{\theta \in \Theta} \sum_{t \in [T]} \alpha_t(\theta) \left(\beta \cdot W_1 \left(P_\theta(D), P_\theta(D_t)\right) + |L(\theta, D_t) - s_t(\theta)|\right)$$

– Depends on smoothness and approximation ability of surrogate functions

– Best achievable

$$\tilde{O}\left(\max_{\theta \in \Theta} \sum_{t \in [T]} \alpha_t(\theta) W_1(P_\theta(D), P_\theta(D_t))\right)$$

– Not necessarily better than pruned HP space for **fixed** weights

– Can match best possible

$$\tilde{O}\left(\max_{\theta \in \Theta} \min_{t \in [T]} W_1(P_\theta(D), P_\theta(D_t))\right)$$

if and only if **weights are adaptive and set appropriately**

$$\alpha_t(\theta) = \begin{cases} 1, & t = \arg\min_{j \in [T]} W_1(P_\theta(D), P_\theta(D_j)) \\ 0, & \text{o.w.} \end{cases}$$

Conclusion

Novel theoretical framework for warm-started few-shot HPO

- Allows understanding of existing meta-learning schemes
- Produces novel insights in terms of the domain-gap and comparison of existing schemes

Role of meta-features

Effect of multi-fidelity evaluation

New warm-started HPO schemes to approach the best possible optimality gap bounds

Thank you

Parikshit Ram
IBM Research

—

Parikshit.Ram@ibm.com

© Copyright IBM Corporation 2022. All rights reserved. The information contained in these materials is provided for informational purposes only, and is provided AS IS without warranty of any kind, express or implied. Any statement of direction represents IBM's current intent, is subject to change or withdrawal, and represent only goals and objectives. IBM, the IBM logo, and ibm.com are trademarks of IBM Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available at Copyright and trademark information.