PED-ANOVA: Efficiently Quantifying Hyperparameter Importance in Arbitrary Subspaces

Shuhei Watanabe , **Archit Bansal** and **Frank Hutter**
Department of Computer Science, University of Freiburg, Germany
{watanabs,bansala,fh}@cs.uni-freiburg.de

Abstract

The recent rise in popularity of Hyperparameter Optimization (HPO) for deep learning has highlighted the role that good hyperparameter (HP) space design can play in training strong models. In turn, designing a good HP space is critically dependent on understanding the role of different HPs. This motivates research on HP Importance (HPI), e.g., with the popular method of functional ANOVA (f-ANOVA). However, the original f-ANOVA formulation is inapplicable to the subspaces most relevant to algorithm designers, such as those defined by top performance. To overcome this issue, we derive a novel formulation of f-ANOVA for arbitrary subspaces and propose an algorithm that uses Pearson divergence (PED) to enable a closed-form calculation of HPI. We demonstrate that this new algorithm, dubbed *PED-ANOVA*, is able to successfully identify important HPs in different subspaces while also being extremely computationally efficient. See https://arxiv.org/abs/2304.10255 for the latest version with Appendix.

1 Introduction

Following on the heels of widespread adoption of deep learning models in various industries and areas of research, Hyperparameter Optimization (HPO) [Bergstra and Bengio, 2012; Snoek *et al.*, 2012; Bergstra *et al.*, 2011; Lindauer *et al.*, 2022; Watanabe, 2023] for deep learning has gained increasing prominence as the path forward for making deep learning more accessible and robust. In particular, recent research has highlighted the role that good hyperparameter (HP) space design can play in training strong models [Chen *et al.*, 2018; Melis *et al.*, 2018; Henderson *et al.*, 2018]. In practice, while a large search space is necessary to find high-performance models [Zimmer *et al.*, 2021], a reduced search space that retains only important HPs is essential for efficiently finding them [Perrone *et al.*, 2019]. Therefore, it is crucial to understand the role that different HPs play in a search space.

This is the driving force behind previous research into the quantification of HP Importance (HPI) [Hutter *et al.*, 2014; Biedenkapp *et al.*, 2017], which still remains a largely understudied section of HPO research. Several HPO frameworks [Biedenkapp *et al.*, 2018; Akiba *et al.*, 2019; Sass *et al.*, 2022] have previously utilized functional ANOVA (f-ANOVA) [Hooker, 2007; Hutter *et al.*, 2014] to provide a better interpretation of the role of different HPs, but the original f-ANOVA formulation is not very practical for the interpretation of specific subspaces of a search space. Such subspaces are often of particular interest to algorithm developers due to various properties, for example, the "local space" visualized in Figure 1 could represent a region of high performance. Nevertheless, prior works [Hutter *et al.*, 2014; Biedenkapp *et al.*, 2018] have attempted to overcome this and quantify HPI in specific subspaces using f-ANOVA. However, since their formulation did not constrain the calculations to subspaces of high interest, we argue that the results are biased towards unimportant subspaces. At the same time, obtaining an unbiased quantification of HPI in specific subspaces of interest, which we refer to as *local* spaces in contrast to the full *global* space, is mathematically non-trivial.

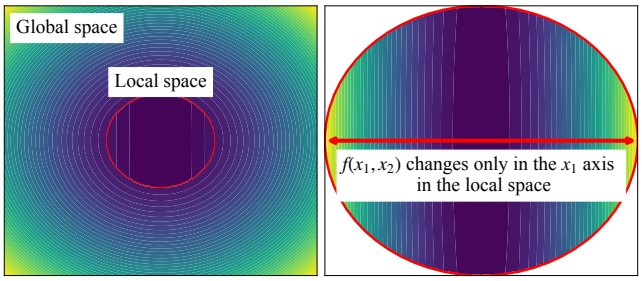
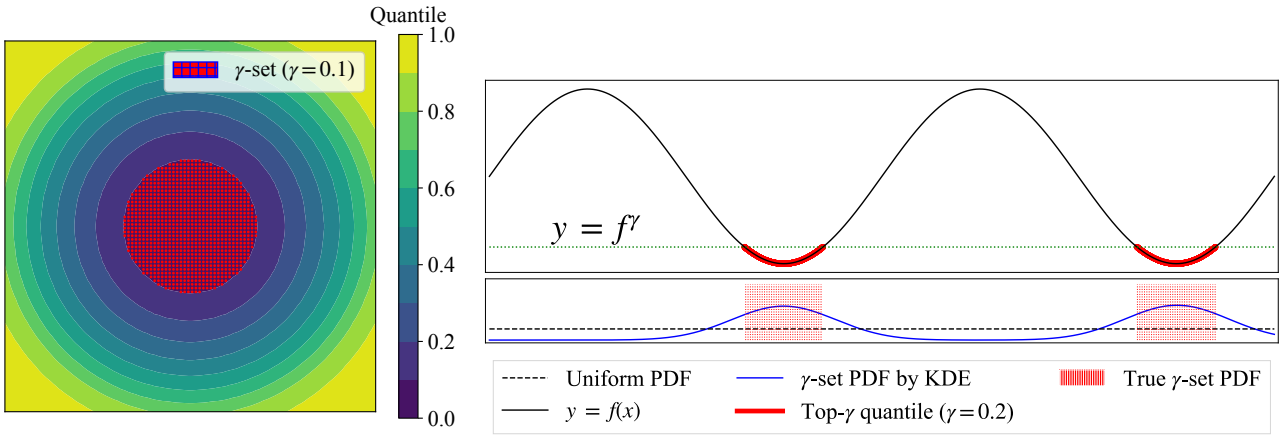


Figure 1: An example where the trend of HPI changes in the global and local space (top-10%). The horizontal axis is the x_1-axis, the vertical axis is the x_2-axis, and $f(x_1, x_2)$ is the objective function to analyze (darker is better). **Left**: the normalized contour plot of $f(x_1, x_2)$ in global space; both x_1 and x_2 appear to variate $f(x_1, x_2)$ equally. The red circle is the promising domain we would like to explore. **Right**: the normalized contour plot of $f(x_1, x_2)$ in the local space (the red circle in global space); $f(x_1, x_2)$ variates only by x_1, and thus x_1 is more important. Such a trend cannot be captured by existing methods.

To overcome this issue, we first formally define local HPI as HPI in a local space and we derive a novel formulation of f-ANOVA to compute local HPI for arbitrary local spaces. Still, our formulation would require Monte-Carlo sampling



(a) The true γ-set of $f(x_1, x_2) = x_1^2 + x_2^2$

(b) An empirical γ-set PDF using KDE of $f(x) = \sin x$

Figure 2: Conceptual visualizations of the γ-set and the γ-set PDF. **Left**: the true γ-set in a 2D example. Darker is better in the figure and we colored the top-10% in red, which is the γ-set $\mathcal{X}^\gamma$ ($\gamma = 0.1$) in this example. **Right**: the true γ-set and the γ-set PDF in a 1D example. The green dotted line shows the γ-quantile value f^γ, which achieves the top-20% in this example, and the red solid lines are the γ-set $\mathcal{X}^\gamma$ ($\gamma = 0.2$). The red dotted spaces below show the true γ-set PDF $p(\boldsymbol{x}|\mathcal{X}^\gamma)$, but since we do not have the analytical form in practice, this PDF is estimated by KDE (the blue solid line).

in general and it is computationally intractable. Therefore, we show that local HPI is tractable without a Monte-Carlo sampling under some constraints and propose an algorithm that uses Pearson divergence (PED, [Pearson, 1900]) to enable a closed-form computation of HPI. In a series of experiments, we first verify that our algorithm correctly provides global and local HPI in a toy function. Then we demonstrate that our algorithm takes only less than a second for 10^5 data points while the prior f-ANOVA [Hutter *et al.*, 2014] would take more than a week.

To provide a solid picture of how to use our method, we perform analysis on JAHS-Bench-201 [Bansal *et al.*, 2022], which has one of the largest search spaces among HPO benchmarks. In the analysis, we find that it is suboptimal to design a search space relying only on global HPI because we potentially miss important HPs in a local space if the global HPI of these HPs are dominated by the most important HP. We demonstrate that local HPI plays a crucial role to avoid this issue. Furthermore, our method has several other possible applications such as (1) post-hoc analysis for HPO, (2) adaptive (e.g., meta-learned) search space reductions for faster HPO, and (3) exploratory data analysis. We discuss these in more detail in Appendix E, along with the advantages and limitations of our method.

In summary, the contributions of this paper are to:

1. reformulate local HPI mathematically and derive the general formula of local HPI,
2. provide a closed-form calculation for local HPI using PED that handles even 10^8 data points in a minute, and
3. benchmark performance compared with the original f-ANOVA.

To facilitate reproducibility, our implementation is available at https://github.com/nabenabe0928/local-anova/.

2 Background & Related Work

2.1 Preliminaries

Throughout this paper, we use the following terms:

1. **γ-quantile value** f^γ: The function value $f^\gamma \in \mathbb{R}$ that achieves the top-γ quantile with respect to the objective function $f : \mathcal{X} \to \mathbb{R}$ to analyze in the global space $\mathcal{X}$,
2. **γ-set** $\mathcal{X}^\gamma$: A set of configurations $\mathcal{X}^\gamma$ that achieves the top-γ quantile in the global space $\mathcal{X}$, and
3. **Marginal γ-set PDF** $p_d(x_d|\mathcal{X}^\gamma)$: The marginal PDF of the γ-set PDF $p(\boldsymbol{x}|\mathcal{X}^\gamma)$:

$$p_d(x_d|\mathcal{X}^\gamma) := \int_{\boldsymbol{x}_{-d} \in \mathcal{X}_{-d}} p(\boldsymbol{x}|\mathcal{X}^\gamma) d\boldsymbol{x}_{-d}. \quad (1)$$

We provide the formal definitions in Appendix B and the conceptual visualizations in Figure 2. Note that $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_D$ is the search space, $\mathcal{X}_d \subseteq \mathbb{R}$ for $d \in [D] := \{1, \dots, D\}$ is the domain of the d-th HP, $\boldsymbol{x}_s \sim \mathcal{X}_s \subseteq \mathbb{R}^{|s|}$ denotes $\boldsymbol{x}_s$ is sampled from the uniform distribution on $\mathcal{X}_s$ where $s \subseteq [D]$, and $\mathcal{X}_{-d} \subseteq \mathbb{R}^{D-1}$ is $\mathcal{X}$ without the d-th dimension. Furthermore, we consistently denote the PDF of the uniform distribution as *uniform PDF* and follow the assumptions stated in Appendix C.1.

2.2 f-ANOVA

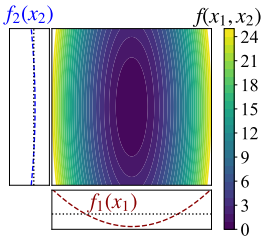
In this section, we describe f-ANOVA for one-dimensional effects and refer to more details about the general version in Appendix A.3. Suppose we would like to quantify HPI of a function $f(\boldsymbol{x})$ defined on $\mathcal{X}$, then global HPI [Hooker, 2007] requires (see Figure 3 for the intuition):

1. **Global mean**:

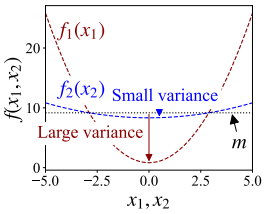
$$m := \mathbb{E}_{\boldsymbol{x} \sim \mathcal{X}}[f(\boldsymbol{x})], \quad (2)$$

2. **Marginal mean**:

$$f_d(x_d) := \mathbb{E}_{\boldsymbol{x}_{-d} \sim \mathcal{X}_{-d}}[f(\boldsymbol{x}|x_d)], \quad (3)$$



(a) The contour plot of $f(x_1, x_2) = x_1^2 + x_2^2/10$ and the marginal means for each dimension.



(b) The conceptual visualization of global HPI on $f(x_1, x_2) = x_1^2 + x_2^2/10$.

Figure 3: The conceptual visualization of global HPI on $f(x_1, x_2) = x_1^2 + x_2^2/10$. **Left**: the landscape of $f(x_1, x_2)$ (darker is better) and the marginal means $f_1(x_1)$ and $f_2(x_2)$. The color does not change a lot in the vertical direction (the x_2-axis) while it does in the horizontal direction (the x_1-axis). **Right**: the marginal means and the mean value in the same plane. As the marginal mean f_1 (the red dashed line) has a large variance, x_1 is more important. On the other hand, as the marginal mean f_2 (the blue dashed line) has a small variance, x_2 is less important.

3. **Marginal variance**:

$$v_d \coloneqq \mathbb{E}_{x_d \sim \mathcal{X}_d}[(f_d(x_d) - m)^2]. \tag{4}$$

Note that $f(\boldsymbol{x}|x_d)$ implies that we fix the d-th HP of $\boldsymbol{x}$ to x_d. When we denote the global variance as v_0, the ratio v_d/v_0 is the *global HPI* of the d-th HP and in essence, the magnitude of the marginal variance represents the relative importance.

2.3 Local f-ANOVA in Prior Works

To the best of our knowledge, there are two papers that mention *local HPI* (and both use f-ANOVA). Hutter *et al.* [2014] mentioned local HPI can be quantified by taking:

$$g(\boldsymbol{x}) \coloneqq \min(f(\boldsymbol{x}), f^\gamma); \tag{5}$$

however, since this measure is biased depending on the global space design as discussed in Appendix B.4.1, we need to consider the integral only over a local space as stated in Section 3.1. Biedenkapp *et al.* [2018] proposed the following HPI measure:

$$\begin{aligned} m_d &\coloneqq \mathbb{E}_{x_d \sim \mathcal{X}_d}[f(\boldsymbol{x}|\boldsymbol{x}^{\text{opt}}_{-d})], \\ v_d &\coloneqq \mathbb{E}_{x_d \sim X_d}[(f(\boldsymbol{x}|\boldsymbol{x}^{\text{opt}}_{-d}) - m_d)^2], \end{aligned} \tag{6}$$

where $\boldsymbol{x}^{\text{opt}} \in \mathbb{R}^D$ is the optimized setting and $\boldsymbol{x}^{\text{opt}}_{-d} \in \mathbb{R}^{D-1}$ is $\boldsymbol{x}^{\text{opt}}$ without the d-th dimension. The authors mention that this measure is a local HPI measure; however, this measure is also not a local HPI measure in our definition and we show that this is the case using a toy example in Appendix B.4.2.

3 Local f-ANOVA Using Pearson Divergence

In this section, we first provide the definition of local HPI and describe how to define a local space. For simplicity, we name this local space definition as *Lebesgue split* [1]. Then we introduce fast algorithm using PED between two KDEs to compute local HPI and benchmark the speed of the algorithm compared to the f-ANOVA implementation based on random forests [Hutter *et al.*, 2014]. Notice that since higher orders of HPI require exponential amounts of computations and usually lack interpretability, our discussion does not focus on higher orders; however, we derive the formula for higher orders and show them in Eqs. (47), (52) in Appendix C. The theoretical details for this section are available in Appendix B.

[1] The name comes from the fact that we define a local space by a function value as in the Lebesgue integral in contrast to the definition of a local space by bounds for each dimension, which we name *Riemann split*.

3.1 Local Hyperparameter Importance

In this section, we assume that we have a set of (sorted) observations $\mathcal{D} \coloneqq \{(\boldsymbol{x}_n, f(\boldsymbol{x}_n))\}_{n=1}^N$ such that $f(\boldsymbol{x}_1) \leq \cdots \leq f(\boldsymbol{x}_N)$. Then, the top-$\gamma$-quantile observations are $\mathcal{D}^\gamma = \{(\boldsymbol{x}_n, f(\boldsymbol{x}_n))\}_{n=1}^{\lceil \gamma N \rceil}$ and the γ-quantile value is $f^\gamma \coloneqq f(\boldsymbol{x}_{\lceil \gamma N \rceil})$.

Local Space Defined by Lebesgue Split

To begin with, we formally define local HPI:

Definition 1 (Local HPI) *Given a subspace $\mathcal{X}^\star \subseteq \mathcal{X}$, local* HPI *is global* HPI v_d/v_0 *in the subspace $\mathcal{X}^\star$.*

Recall that v_d is the marginal variance of the d-th HP and v_0 is the global variance. Based on Definition 1, the prior works on making f-ANOVA local discussed in Section 2.3 are not local HPI measures; see Appendix B.4 for more details. As our local HPI obviously depends on the choice of $\mathcal{X}^\star$, local HPI is a very general concept; therefore, we focus on the so-called *Lebesgue split* to specify a local space in this paper. In the Lebesgue split, we obtain a local space $\mathcal{X}^\star$ as follows:

1. Fix a threshold $f^\star$ (we use the γ-quantile value f^γ in this paper instead), and
2. Obtain the sublevel set $\mathcal{X}^\star \coloneqq \{\boldsymbol{x} \in \mathcal{X} | f(\boldsymbol{x}) \leq f^\star\}$ based on $f^\star$ ($\mathcal{X}^\star$ becomes the γ-set $\mathcal{X}^\gamma$ when $f^\star = f^\gamma$).

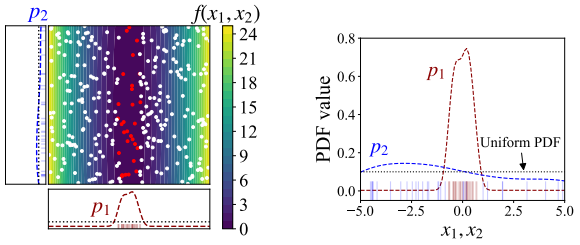
Recall that the definitions of the γ-quantile value and the γ-set are available in Section 2.1. More intuitively, the red domains in Figure 2 are the local space of each example. In this paper, we use $f(\boldsymbol{x}_{\lceil \gamma N \rceil})$ as f^γ. The advantages of the Lebesgue split are to:

1. require only one parameter $f^\star$ while the Riemann split, which we split along each dimension by specifying bounds, requires at least $2 \times D$ parameters,
2. be able to focus on the analysis in promising domains where we are interested, and
3. be able to remove the sampling bias caused by a non-uniform sampler when using the formula of local HPI.

We further discuss the strengths and drawbacks of the Lebesgue split compared to the Riemann split in Appendix B.5.

Formula of Local Hyperparameter Importance

Now we discuss the computation of local HPI. In Eqs. (2)–(4), we take the expectation over the uniform distribution of the global space $\mathcal{X}$. In the same vein, it is natural to consider the expectation over the uniform distribution of the local space $\mathcal{X}^\gamma$ for local HPI as well. Although the computation is not obvious, we can compute the expectation of a measurable



(a) The contour plot of $f(x_1, x_2)$ and the marginal γ'-set PDFs.

(b) The conceptual visualization of global HPI by PED.

Figure 4: The conceptual visualization of global HPI using PED on $f(x_1, x_2) = x_1^2 + x_2^2/100$. As we consider global HPI in this example, the γ-set PDF is the uniform PDF. **Left**: the landscape of $f(x_1, x_2)$ (darker is better) and the marginal γ'-set PDFs $p_1(x_1|\mathcal{D}^{\gamma'})$ ($= p_1$, the red dashed line) and $p_2(x_2|\mathcal{D}^{\gamma'})$ ($= p_2$, the blue dashed line). The dots represent observations (datasets) and the red dots achieve the top-γ' quantile. p_1, p_2 are estimated by Eq. (15) using the red dots. **Right**: the marginal γ'-set PDFs in the same plane. While p_1 (the red dashed line) sharply peaks at the center, p_2 (blue dashed line) is close to the uniform PDF. This implies that the latter is closer to the uniform PDF, and thus x_2 is less important.

function $f(\boldsymbol{x})$ over the local space $\mathcal{X}^\gamma$ if we use the following trick:

$$\frac{1}{\gamma}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}[f(\boldsymbol{x})\mathbb{I}[\boldsymbol{x}\in\mathcal{X}^\gamma]], \tag{7}$$

where $\gamma = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}[\mathbb{I}[\boldsymbol{x}\in\mathcal{X}^\gamma]]$ is a normalization constant. Recall that $\boldsymbol{x}\in\mathcal{X}^\gamma$ and $f(\boldsymbol{x})\le f^\gamma$ are equivalent. Similarly, the marginal mean of $f(\boldsymbol{x})$ is computed as follows:

$$\frac{1}{V_d^\gamma(x_d)}\mathbb{E}_{\boldsymbol{x}_{-d}\sim\mathcal{X}_{-d}}[f(\boldsymbol{x}|x_d)\mathbb{I}[\boldsymbol{x}\in\mathcal{X}^\gamma|x_d]], \tag{8}$$

where $V_d^\gamma(x_d) \coloneqq \mathbb{E}_{\boldsymbol{x}_{-d}\sim\mathcal{X}_{-d}}[\mathbb{I}[\boldsymbol{x}\in\mathcal{X}^\gamma|x_d]]$ is a normalization constant. Then local HPI is generally computed as follows:

1. **Local mean**:

$$m^\gamma \coloneqq \frac{1}{\gamma}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}[f(\boldsymbol{x})\mathbb{I}[\boldsymbol{x}\in\mathcal{X}^\gamma]], \tag{9}$$

2. **Local marginal mean**:

$$f_d^\gamma(x_d) \coloneqq \frac{1}{V_d^\gamma(x_d)}\mathbb{E}_{\boldsymbol{x}_{-d}\sim\mathcal{X}_{-d}}[f(\boldsymbol{x}|x_d)\mathbb{I}[\boldsymbol{x}\in\mathcal{X}^\gamma|x_d]], \tag{10}$$

3. **Local marginal variance**:

$$v_d^\gamma \coloneqq \mathbb{E}_{x_d\sim V_d^\gamma}[(f_d^\gamma(x_d) - m^\gamma)^2]. \tag{11}$$

Note that $x_d \sim V_d^\gamma$ implies that x_d is sampled from the distribution of the PDF $V_d^\gamma(x_d)/Z$ where $Z\in\mathbb{R}_+$ is the normalization constant. As the series of computations requires a Monte-Carlo sampling in a $D-1$ dimensional space, the time complexity incurs the curse of dimensionality. In the next section, we introduce fast algorithm to compute local HPI in exchange for the scale ignorance.

Algorithm 1 Local PED-ANOVA

$\mathcal{D} = \{(\boldsymbol{x}_n, f(\boldsymbol{x}_n))\}_{n=1}^N$ (Dataset to analyze), γ, γ' (User-defined quantiles of top domains)
▷ See Appendices E.2, E.3 for practical usages
Sort $\mathcal{D}$ in ascending order by f
▷ $|\mathcal{D}^\gamma| \ge 2$ and $|\mathcal{D}^{\gamma'}| \ge 2$ must hold
Pick the top-γ and -γ' quantile observations $\mathcal{D}^\gamma, \mathcal{D}^{\gamma'}$
for $d = 1, \dots, D$ **do**
    Count occurrences of unique values $c_d^{(n)}$
    Build KDEs $p_d(\cdot|\mathcal{D}^\gamma), p_d(\cdot|\mathcal{D}^{\gamma'})$ by Eq. (15)
    Compute v_d^γ by Eq. (16)
return $\{v_d^\gamma\}_{d=1}^D$

3.2 Fast Algorithm by Pearson Divergence

If we analyze the binary function $b(\boldsymbol{x}|\mathcal{X}^{\gamma'}) \coloneqq \mathbb{I}[\boldsymbol{x}\in\mathcal{X}^{\gamma'}] = \mathbb{I}[f(\boldsymbol{x})\le f^{\gamma'}]$ instead of $f(\boldsymbol{x})$, HPI can be efficiently computed where $\gamma'(<\gamma)$ is another quantile to define the binary function in the local space $\mathcal{X}^\gamma$. First, we prove the following theorem:

Theorem 1 *Given the binary function $b(\boldsymbol{x}|\mathcal{X}^{\gamma'})$ and the γ'- and γ-set* PDFs *$p(\boldsymbol{x}|\mathcal{X}^{\gamma'}), p(\boldsymbol{x}|\mathcal{X}^\gamma)$ where $\gamma' < \gamma$, the local marginal variance of each dimension $d\in[D]$ is:*

$$v_d^\gamma = \left(\frac{\gamma'}{\gamma}\right)^2 D_{\mathrm{PE}}(p_d(\cdot|\mathcal{X}^{\gamma'})\|p_d(\cdot|\mathcal{X}^\gamma)). \tag{12}$$

The proof is provided in Appendix C.3 and higher orders of HPI can be computed by Eq. (52) in Appendix C.3. Note that PED between the PDFs p, q defined on $\mathcal{X}_d$ is computed as:

$$D_{\mathrm{PE}}(p\|q) \coloneqq \mathbb{E}_{x_d\sim q(x_d)}\left[\left(\frac{p(x_d)}{q(x_d)} - 1\right)^2\right]. \tag{13}$$

As we do not have the ground truth of the marginal γ'- and γ-set PDFs, we replace them with KDEs. The tricks of this computation are that (1) the marginal γ-set PDF can be easily estimated by (1D) KDE as follows and (2) we only need to take the average in 1D space:

$$p_d(x_d|\mathcal{D}^\gamma) \coloneqq \frac{1}{\lceil\gamma N\rceil}\sum_{n=1}^{\lceil\gamma N\rceil} k(x_{n,d}, x_d). \tag{14}$$

Note that $x_{n,d}\in\mathcal{X}_d$ is the d-th dimension of $\boldsymbol{x}_n$ and k is a kernel function. Although the query of this function still requires $O(N)$, the time complexity scales down to $O(n_d)$ where $n_d\in\mathbb{Z}_+$ is the number of unique values in the d-th HP if we use the following compression:

$$p_d(x_d|\mathcal{D}^\gamma) = \frac{1}{\lceil\gamma N\rceil}\sum_{n=1}^{n_d} c_d^{(n)} k(x_d^{(n)}, x_d) \tag{15}$$

where $x_d^{(n)}$ is the n-th unique value in the d-th HP and $c_d^{(n)}$ is the occurrences of this value in $\mathcal{D}^\gamma$. Note that we discretize a continuous HP $x_d\in[L,R] (L<R)$ (if exists) as $x_d\in\{L + n(R-L)/(n_d-1)\}_{n=0}^{n_d-1}$ to apply Eq. (15)

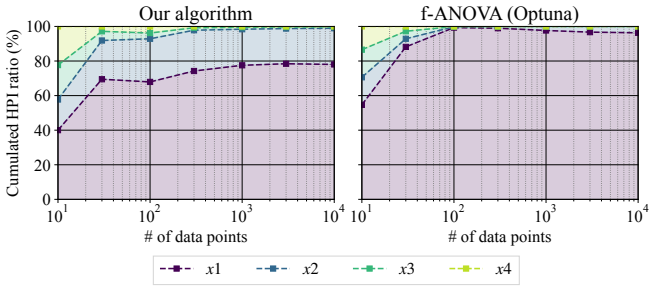


Figure 5: The comparison of global HPI between our algorithm (**Left**) and Optuna f-ANOVA (**Right**). HPI is averaged over 10 runs with different random seeds. The horizontal axis shows the number of data points N and the vertical axis shows the cumulative HPI ratio. HPI ratio is computed by $v_d / \sum_{d'=1}^{D} v_{d'}$ and the weak color band between each plot shows the HPI ratio of each HP.

and the discretization error of marginal variances is bounded by $O(\frac{1}{n_d})$ under some assumptions. Since we can avoid Monte-Carlo samplings with the discretization and the total time complexity is reduced to $O(N + n_d^2)$, this is a trade-off; see Proposition 2 in Appendix C.4 for more details. Hence Eq. (11) is approximated as the following closed-form:

$$v_d^\gamma \simeq \left(\frac{\gamma'}{\gamma}\right)^2 \sum_{n=1}^{n_d} \frac{p_d(x_d^{(n)}|\mathcal{D}^\gamma)}{Z} \left(\frac{p_d(x_d^{(n)}|\mathcal{D}^{\gamma'})}{p_d(x_d^{(n)}|\mathcal{D}^\gamma)} - 1\right)^2 \tag{16}$$

where $Z \coloneqq \sum_{n=1}^{n_d} p_d(x_d^{(n)}|\mathcal{D}^\gamma)$ is a normalization constant and the time complexity of this computation is $O(n_d^2)$. Algorithm 1 shows the pseudocode for the local HPI computation. Note that global HPI, whose computation is detailed in Appendix B.2, can be computed by replacing $p_d(\cdot|\mathcal{D}^\gamma)$ with the uniform PDF in the d-th dimension $u_d(x_d)$ (or $p_d(\cdot|\mathcal{D}^\gamma)$ with $\gamma = 1$, i.e. $p_d(\cdot|\mathcal{D})$, as discussed in Appendix E.2 when collecting $\mathcal{D}$ by a non-uniform sampler). Figure 4 presents an example of global HPI with our method on a 2D toy function.

4 Performance Validation

4.1 Setup

In this section, we consistently use the following function:

$$f(x_1, x_2, x_3, x_4) = \sum_{d=1}^{4} w_d(x_d) \times x_d^2 \tag{17}$$

where $x_d \in [-5, 5]$ for all $d \in \{1, 2, 3, 4\}$ and the weights $w_d : \mathbb{R} \to \boldsymbol{W}$ follow:

$$w_d(x) = \begin{cases} W_{d-1} & (|x| \geq 1) \\ W_{d+2 \bmod 4} & (\text{otherwise}) \end{cases}. \tag{18}$$

and $\boldsymbol{W} \coloneqq \{W_d\}_{d=0}^3 = \{5^0, 5^{-1}, 5^{-2}, 5^{-3}\}$. This function has different trends of HPI in global and local spaces. While the order of HPI is x_1, x_2, x_3, x_4 in the global space, it is x_2, x_3, x_4, x_1 in the local space $\{\boldsymbol{x} \in \mathcal{X} | \forall d \in [4], |x_d| < 1\}$.

In this experiment, we discretized the HPs with $n_d = 1001$ and all samples were drawn from the uniform distribution. Furthermore, all experiments were run on the hardware with Intel Core i7-10700 and we used the f-ANOVA implementation with the default parameter setting by Optuna [2]. Note that the Optuna implementation is based on Hutter *et al.* [2014].

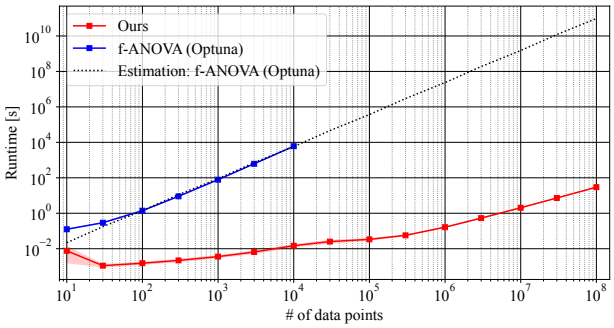


Figure 6: The benchmark of query speed of our method and f-ANOVA with respect to the number of data points N. Each setting was run with 10 different seeds and the weak color bands show the standard error. As f-ANOVA requires much more computation, we estimated the evolution and provided the estimation by the black dotted line.

4.2 Effect of Scale Ignorance in Global HPI

Since PED-ANOVA uses $\mathbb{I}[f(\boldsymbol{x}) \leq f^\star]$ instead of $f(\boldsymbol{x})$, it cannot capture scale information. On the other hand, since our objective is to identify important HPs, we would like to test if PED-ANOVA can identify important HPs. In the experiment, we used $\gamma' = 0.1$. Figure 5 shows the cumulative global HPI ratio of each method. As seen in the figure, while both methods could identify the most important HP x_1, we can see the difference in the HPI of x_2. While PED-ANOVA tells us x_2 has about 20% of contribution to achieve the top-10%, f-ANOVA tells us x_2 has about 3% of contribution. This difference comes from whether we ignore the scale of the objective function or not. Since f-ANOVA considers scale and it magnifies the contribution in the tail of the function, it dilutes the HPI of x_2, which has less weight in the tail. Note that "tail" refers to the domains that cause a lot of variations, yet not critical for the final result, in the objective function f and $|x_d| \geq 1$ is the tail in our case; more details in Appendix B.6. On the other hand, the HPI of x_1 by PED-ANOVA is not strongly biased by the tail due to the scale ignorance nature. This leads to more importance in x_2. Although our method loses scale information, the ignorance of scale allows us to abandon the information from the tail and focus only on the information from the promising domain, which is γ'-set in our case. Furthermore, this remarkable property makes the meaning of HPI, which is how important each HP is to achieve the top-γ' quantile, very clear and practitioners can extract the nuance of each HP for specific local spaces.

4.3 Query Speed

As mentioned previously, one of the benefits of our method is the query speed, and we would like to benchmark how quick our method is in this section. In the experiment, we used $\gamma' = 0.1$. Figure 6 presents the runtime with respect to the number of data points. While f-ANOVA requires more than a

[2] https://github.com/optuna/optuna

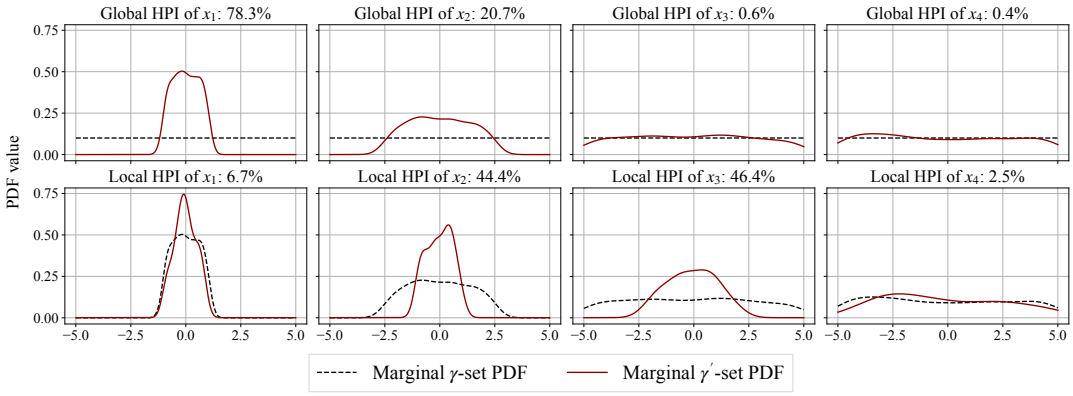


Figure 7: The validation of the local importance measure. The black dashed lines show the marginal γ-set PDFs and the red lines show the marginal γ'-set PDFs for each dimension. The percentage (**HPI ratio**) was computed by $v_d / \sum_{d'=1}^{D} v_{d'}$. **Top row**: the plots of the uniform PDFs ($\gamma = 1$) and the marginal $\gamma' = 0.1$-set PDFs. These are used to compute global HPI. **Bottom row**: the plots of the marginal $\gamma = 0.1$-set PDFs and the marginal $\gamma' = 0.01$-set PDFs. Those are used to compute local HPI in the top-10% domain.

minute with 10^3 data points and more than a week with 10^5 data points, our method provides the results in a minute even with 10^8 data points. In Appendix D.2, we test our method with higher dimensionality to see the number of data points required for convergence.

4.4 Local Importance Measure

Finally, we check if our method can successfully identify important HPs in promising domains. The objective function $f(\boldsymbol{x})$ is designed so that while x_1 is important and x_3 is trivial in the global space, x_1 is less important and x_3 is important in the local space. The goal of this experiment is to check whether our method can provide this insight. In the experiment, we used $N = 10^4$.

Figure 7 shows the results. As discussed already, global HPI could identify the order of HPI appropriately. For local HPI, our method could tell us that x_2, x_3 are the most important HPs in the local space and x_1 is less important as expected. Note that since the $\gamma = 0.1$-set already narrows down the domain of x_2, but not x_3, this dilutes the HPI of x_2 and increases the HPI of x_3. Prior works cannot provide this interpretation as discussed in Appendix B.4.

5 Real-World Usecase by JAHS-Bench-201

5.1 Setup

In order to further verify our proposed algorithm against a real-world application, we applied PED-ANOVA to analyze the search space of JAHS-Bench-201 [Bansal *et al.*, 2022], which is a surrogate benchmark for HPO and has a very large search space in the context of extant HPO benchmarks. We constructed the dataset $\mathcal{D}$ in Algorithm 1 by querying JAHS-Bench-201 for the validation accuracy, i.e. $f(\boldsymbol{x})$, of N lattice points, where $N = 41{,}343{,}750$, generated by discretizing the JAHS-Bench-201 search space (see Table 2 of Appendix D.1). Although JAHS-Bench-201 can be queried for model performance metrics on 3 different image classification datasets, for the sake of brevity, we discuss only the experiments performed on `CIFAR10` here and include the results on the other datasets in Appendix D.3. Due to the computational complexity of f-ANOVA, we could use only 10^4 data points for it, in contrast to PED-ANOVA, and calculated the mean of HPI over 10 independent runs. Since the surrogate models in JAHS-Bench-201 are trained XGBoost models and XGBoost's outputs are deterministic, we query each lattice point only once. In the analysis, we would like to answer the following research questions (RQs):

	HPI ratio (%)					
Hyperparameter	Normal	Global 0.1		Global 0.01		Local
	Original	Ours	Original	Ours	Original	Ours
Learning rate	1.36	**9.11**	**10.20**	6.62	3.59	4.09
Weight decay	0.96	2.19	0.68	2.56	0.31	3.00
Activation function	0.01	0.12	0.21	0.26	0.41	0.40
TrivialAugment	0.00	4.33	**3.83**	**13.22**	**8.27**	**28.33**
Depth multiplier	0.06	0.66	0.58	2.47	0.63	6.90
Width multiplier	1.60	**60.22**	**73.59**	**35.26**	**71.75**	9.07
Operation 1 (Op.1)	**11.86**	**6.65**	3.45	**11.95**	3.81	**13.38**
Operation 2 (Op.2)	4.04	2.36	1.42	5.00	2.51	6.97
Operation 3 (Op.3)	**64.73**	5.63	1.14	5.25	1.73	5.50
Operation 4 (Op.4)	0.09	0.84	0.83	1.62	1.09	2.09
Operation 5 (Op.5)	4.00	2.19	1.04	4.72	1.29	6.76
Operation 6 (Op.6)	**11.29**	5.71	3.02	11.06	**4.61**	**13.52**

Table 1: HPI of `CIFAR10` in JAHS-Bench-201. The ratio of HPI by percentage (**HPI ratio**) computed by $v_d / \sum_{d'=1}^{D} v_{d'}$. The top-3 HPs are bolded. **Cols. 1,3,5** (Original): HPI by f-ANOVA on $g(\boldsymbol{x}) := \min(f(\boldsymbol{x}), f^{\gamma'})$. **Cols. 2,4,6** (Ours): HPI by PED-ANOVA.

RQ1: Does global HPI of our method provide the same important HPs as f-ANOVA with Eq. (5)?

RQ2: Is the scale ignorance necessary for matching the intuition of achieving the top-γ' quantile?

RQ3: Does local HPI help detect potentially important HPs or trivial HPs?

In order to answer RQs, we provide Table 1 with the HPI of each HP in `CIFAR10` of JAHS-Bench-201 and Figure 8 to visualize the $\gamma = 0.1$- and $\gamma = 0.01$-set PDFs as the blue and the red shadows, respectively. Strictly speaking, discrete probability distributions are not PDFs due to discrete space; however, we use the term γ-set PDF for the sake of consistency. We applied f-ANOVA to $f(\boldsymbol{x})$ (`Normal`), $\min(f(\boldsymbol{x}), f^{\gamma'=0.1})$ (`Global 0.1`), and $\min(f(\boldsymbol{x}), f^{\gamma'=0.01})$ (`Global 0.01`), and PED-ANOVA

Figure 8: The distributions of important HPs of `CIFAR10` in JAHS-Bench-201. The red shadows show the $\gamma = 0.01$-set PDFs, the blue shadows show the $\gamma = 0.1$-set PDFs, and the black dashed lines show the uniform PDFs. PED between a black line and a blue shadow is `Global 0.1`, PED between a black line and a red shadow is `Global 0.01`, and PED between a red shadow and a blue shadow is `Local` in Table 1. The titles for each figure show the names of each HP and the details of HPs are available in Appendix D.1. Notice that C1 – C5 correspond to the order of Table 2 and the overlap between the red and the blue shadows looks purple although they are separated shadows.

with $\gamma' = 0.1$ (`Global 0.1`), $\gamma' = 0.01$ (`Global 0.01`), and $\gamma = 0.1, \gamma' = 0.01$ (`Local`). Recall that `Global 0.1` and `Global 0.01` for f-ANOVA are based on Eq. (5) and $f^{\gamma'=0.1}$ means $f_{\lceil |\mathcal{D}|/10 \rceil}$ given a dataset sorted by f_n. Although we used the uniform PDF to compute global HPI in this experiment, practitoners should use $p_d(\cdot|\mathcal{D})(\gamma = 1)$ instead of the uniform PDF for the post-hoc analysis of HPO when using a non-uniform sampler (e.g. Bayesian optimization) to remove sampling bias as discussed in Appendix E.2.

5.2 Analysis & Interpretation

To answer RQ1, we compare the column `(Global 0.1, Ours)` to `(Global 0.1, Original)` and the column `(Global 0.01, Ours)` to `(Global 0.01, Original)` in Table 1. We observe that both PED-ANOVA and f-ANOVA indicated the same top-2 important HPs although the 3$^{\text{rd}}$-best HPs slightly varied. This result further verifies the validation in Section 4.2.

To answer RQ2, we discuss the results of `(Global 0.1, Ours)` and `(Global 0.01, Ours)` in the context of `(Normal, Original)` to assess the impact that the tail of $f(\boldsymbol{x})$ discussed in Section 4.2 has on f-ANOVA. The most important takeaway from this comparison is the misclassification of `Op.3` as the most important and of `TrivialAugment` as the least important HP to optimize over by the original f-ANOVA in the global setting. As can be verified by looking at the γ-set PDFs, even for $\gamma = 0.01$, `Op.3`'s values are distributed very evenly even when `TrivialAugment` and `Width multiplier` have already shown convergence. This clearly indicates that `Op.3` is not very important to optimize for achieving the top-1% performance and may or may not become relevant in even higher quantile regimes. At the same time, both columns' values agree on the importance of `Op.1` and `Op.6`. Therefore, to answer RQ2, scale invariance indeed helps to successfully identify HPI for HPs that would have been misclassified by the `(Normal, Original)` setting.

Finally, for RQ3, we compare the column `(Global 0.01, Ours)` to `(Local, Ours)`. We observe that the HPI of `Width multiplier` drops sharply from the `Global 0.01` setting to the `Local` setting. Simultaneously, the HPI of `TrivialAugment` increases sharply across the same. This suggests that optimizing `Width multiplier` is no longer important when moving from the top-10% to the top-1% performance but optimizing `TrivialAugment` is very important. The reason behind this change becomes clear when we observe the change in γ-set PDFs of the two HPs in Figure 8. Both the γ-set PDFs for `Width multiplier` are sharply peaked at 16, indicating that no further optimization is needed on `Width multiplier`. However, the γ-set PDFs for `TrivialAugment` only start peaking at the value `True` for the $\gamma = 0.01$-set PDF. This clearly demonstrates that local HPI is necessary for deriving the correct interpretation in the top-γ' quantiles, since `(Local, Ours)` successfully identifies the relative importance of optimizing the two HPs. Last but not least, if both global and local HPI with wished quantiles γ, γ' exhibits low values, removing such HPs, e.g. `Activation function`, is expected to have a less negative impact although it is insecure to remove HPs, e.g. `Op.1`, only by looking at global HPI.

6 Conclusions

In this paper, we reformulated f-ANOVA for local HPI and introduced the fast algorithm to compute local HPI by PED. In the series of experiments on a toy function, we confirmed that our method can quantify both global and local HPI appropriately, and efficiently compute HPI in a second with 10^5 data points while the prior work takes several days. In the analysis of JAHS-Bench-201, we provided a concrete example of how to use our method on benchmark datasets and showed that only using global HPI could be misleading. Due to the space limit, we defer a discussion of practical usecases and limitations of our method to Appendix E. Our implementation is available at https://github.com/nabenabe0928/local-anova/.

Acknowledgments

The authors appreciate the valuable contributions of the anonymous reviewers. Robert Bosch GmbH is acknowledged for financial support. The authors also acknowledge funding by European Research Council (ERC) Consolidator Grant "Deep Learning 2.0" (grant no. 101045765). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the ERC. Neither the European Union nor the ERC can be held responsible for them.

Funded by
the European Union

References

[Akiba *et al.*, 2019] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *International Conference on Knowledge Discovery & Data Mining*, 2019.

[Bansal *et al.*, 2022] A. Bansal, D. Stoll, M. Janowski, A. Zela, and F. Hutter. JAHS-Bench-201: A foundation for research on joint architecture and hyperparameter search. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[Bergstra and Bengio, 2012] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 2012.

[Bergstra *et al.*, 2011] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, 2011.

[Biedenkapp *et al.*, 2017] A. Biedenkapp, M. Lindauer, K. Eggensperger, F. Hutter, C. Fawcett, and H. Hoos. Efficient parameter importance analysis via ablation with surrogates. In *Association for the Advancement of Artificial Intelligence*, 2017.

[Biedenkapp *et al.*, 2018] A. Biedenkapp, J. Marben, M. Lindauer, and F. Hutter. CAVE: Configuration assessment, visualization and evaluation. In *International Conference on Learning and Intelligent Optimization*, 2018.

[Chen *et al.*, 2018] Y. Chen, A. Huang, Z. Wang, I. Antonoglou, J. Schrittwieser, D. Silver, and N. de Freitas. Bayesian optimization in AlphaGo. *arXiv:1812.06855*, 2018.

[Henderson *et al.*, 2018] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. In *Association for the Advancement of Artificial Intelligence*, 2018.

[Hooker, 2007] G. Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16, 2007.

[Hutter *et al.*, 2014] F. Hutter, H. Hoos, and K. Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *International Conference on Machine Learning*, 2014.

[Lindauer *et al.*, 2022] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhkopf, R. Sass, and F. Hutter. SMAC3: A versatile bayesian optimization package for Hyperparameter Optimization. *Journal of Machine Learning Research*, 23, 2022.

[Melis *et al.*, 2018] G. Melis, C. Dyer, and P. Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*, 2018.

[Pearson, 1900] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine and Journal of Science*, 50, 1900.

[Perrone *et al.*, 2019] V. Perrone, H. Shen, MW. Seeger, C. Archambeau, and R. Jenatton. Learning search spaces for Bayesian optimization: Another view of hyperparameter transfer learning. *Advances in Neural Information Processing Systems*, 2019.

[Sass *et al.*, 2022] R. Sass, E. Bergman, A. Biedenkapp, F. Hutter, and M. Lindauer. DeepCAVE: An interactive analysis tool for automated machine learning. *arXiv:2206.03493*, 2022.

[Snoek *et al.*, 2012] J. Snoek, H. Larochelle, and R. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, 2012.

[Watanabe, 2023] S. Watanabe. Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv:2304.11127*, 2023.

[Zimmer *et al.*, 2021] L. Zimmer, M. Lindauer, and F. Hutter. Auto-Pytorch: Multi-fidelity metalearning for efficient and robust AutoDL. *Transactions on Pattern Analysis and Machine Intelligence*, 43, 2021.