A APPENDIX

We add the proof of Theorem 1 and additional numerical experiments here.

A.1 PRELIMINARIES FROM OPTIMAL TRANSPORT THEORY

Definition 2. Suppose X is a metric space equipped with the metric $d(\boldsymbol{x}, \boldsymbol{y})$, and μ and ν are two probability measures on X. The *Wasserstein distance* (as known as the *Kantorovich–Rubinstein metric*) $d_{W^d}(\mu, \nu)$ between to probability measures μ and ν for the metric function $d(\boldsymbol{x}, \boldsymbol{y})$ is defined to be

$$d_{W^d}(\mu, \nu) = \inf_{\pi \in \Pi(X \times X)} \int_{X \times X} d(\boldsymbol{x}, \boldsymbol{y}) \, d\pi(\boldsymbol{x}, \boldsymbol{y}),$$

where $\Pi(X \times X)$ is the collection of all probability measure on $X \times X$ such that

$$\pi(A \times X) = \mu(A), \quad \pi(X \times B) = \nu(B)$$

for all measurable sets $A, B \subset X$.

For the analysis of the adaptive algorithm in this work, we consider the metric $d_M(\boldsymbol{x}, \boldsymbol{y})$ induced by the Euclidean metric $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$

$$d_M(\boldsymbol{x}, \boldsymbol{y}) = \min\{M, d(\boldsymbol{x}, \boldsymbol{y})\}, \quad \boldsymbol{x}, \boldsymbol{y} \in X.$$

Then the metric $d_M(\boldsymbol{x}, \boldsymbol{y})$ is always bounded by M (reachable, namely $\|d_M\|_\infty = M$). We denote the Wasserstein distance for $d_M(\boldsymbol{x}, \boldsymbol{y})$ by $d_{W^M}(\cdot, \cdot)$.

According to the optimal transport theory, the Wasserstein distance can be described by its dual form (see e.g. Villani (2003), Theorem 1.14 and Remark 1.15 on Page 34).

Theorem 3 (Kantorovich-Rubinstein theorem)**.** *Let X be a Polish space and let d be a lower semi-continuous metric on X. Let $\|\cdot\|_{Lip}$ denote the Lipschitz norm of a function defined as*

$$\|\phi\|_{Lip} = \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{|\phi(\boldsymbol{x}) - \phi(\boldsymbol{y})|}{d(\boldsymbol{x}, \boldsymbol{y})}.$$

Then

$$d_{W^M}(\mu, \nu) = \sup \Big\{ \int_X \phi(\boldsymbol{x}) \, d(\mu - \nu)(\boldsymbol{x}) \Big| \, 0 \leq \phi(\boldsymbol{x}) \leq \|d_M\|_\infty = M, \text{ and } \|\phi\|_{Lip} \leq 1 \Big\}.$$

In this work, we restrict ourselves on a compact domain $X = \Omega \subset \mathbb{R}^D$ of learning, and without loss of generality, we assume the Lebesgue measure of Ω is 1.

A.2 THE FIRST CONVERGENCE THEOREM AND ITS PROOF

Theorem 4. *Let μ be the Lebesgue measure on X, which represents the uniform probability distribution on Ω. In addition, we assume Assumption A1 holds.*

Then the optimal value of the min-max problem equation 5 is 0. Moreover, there is a sequence $\{u_n\}_{n=1}^\infty$ of functions with $r(u_n) \neq 0$ for all n, such that it is an optimization sequence of problem equation 5, namely,

$$\lim_{n \to \infty} \mathcal{J}(u_n, p_n) = 0. \tag{15}$$

for some sequence of functions $\{p_n\}_{n=1}^\infty \subset V$. Meanwhile, this optimization sequence has the following two properties:

1. *The residual sequence $\{r(u_n)\}_{n=1}^\infty$ of $\{u_n\}_{n=1}^\infty$ converges to 0 in $L^2(d\mu)$.*
2. *The renormalized squared residual distributions*

$$d\nu_n \triangleq \frac{r^2(u_n)}{\int_\Omega r^2(u_n(\boldsymbol{x})) \, d\boldsymbol{x}} \, d\mu \tag{16}$$

converge to the uniform distribution μ in the Wasserstein distance d_{W^M}.

Proof. Consider a minimizing sequence $u_n, n = 1, 2, \dots$ of

$$\inf_u \int_\Omega r^2(u(\boldsymbol{x}))\, d\boldsymbol{x}, \tag{17}$$

where without loss of generality, we can assume that $\int_\Omega r^2(u_n(\boldsymbol{x}))\, d\boldsymbol{x} \leq \frac{1}{n}$.

Now

$$\begin{aligned}
&\sup_{\substack{\|p\|_{\text{Lip}}\leq 1\\ 0\leq p\leq M}} \mathcal{J}(u_n, p)\\
&= \sup_{\substack{\|p\|_{\text{Lip}}\leq 1\\ 0\leq p\leq M}} \Big[\int_\Omega r^2(u_n(\boldsymbol{x}))p(\boldsymbol{x})\, d\boldsymbol{x} - \int_\Omega r^2(u_n(\boldsymbol{x}))\, d\boldsymbol{x}\int_\Omega p(\boldsymbol{x})\, d\boldsymbol{x} + \int_\Omega r^2(u_n(\boldsymbol{x}))\, d\boldsymbol{x}\int_\Omega p(\boldsymbol{x})\, d\boldsymbol{x}\Big]\\
&\leq \int_\Omega r^2(u_n(\boldsymbol{x}))\, d\boldsymbol{x}\Big(\sup_{\substack{\|p\|_{\text{Lip}}\leq 1\\ 0\leq p\leq M}} \Big[\int_\Omega p(\boldsymbol{x})\, d\nu_n(\boldsymbol{x}) - \int_\Omega p(\boldsymbol{x})\, d\boldsymbol{x}\Big] + \sup_{\substack{\|p\|_{\text{Lip}}\leq 1\\ 0\leq p\leq M}} \int_\Omega p(\boldsymbol{x})\, d\boldsymbol{x}\Big)\\
&= (d_{W^M}(\nu_n, \mu) + M)\int_\Omega r^2(u_n(\boldsymbol{x}))\, d\boldsymbol{x}.
\end{aligned} \tag{18}$$

By the assumption of the theorem, for each n, we can find a function $\tilde{u}_n(\boldsymbol{x})$ so that the Wasserstein distance $d_{W^M}(\tilde{\nu}_n, \mu) \leq \frac{1}{n}$, where $\tilde{\nu}_n$ is the measure defined as in equation 16 by replacing $u_n(\boldsymbol{x})$ with $\tilde{u}_n(\boldsymbol{x})$. In fact, for each n, we can find, by partition of unity, a sequence of functions in $C_c^\infty(\Omega)$ converging to $\mathbb{1}_\Omega$ in the Sobolev norm of $W^{k,1}$ (See for example Evans (2010)). So we can find a function w_n in $C_c^\infty(\Omega)$, such that $\|w_n(\boldsymbol{x}) - \mathbb{1}_\Omega(\boldsymbol{x})\|_1 \leq \frac{1}{n}$ on Ω. Since r is a surjection, there is some $\tilde{u}_n(\boldsymbol{x})$ so that

$$r^2(\tilde{u}_n) = w_n \int_\Omega r^2(u_n(\boldsymbol{x}))\, d\boldsymbol{x},$$

and

$$\begin{aligned}
\int_\Omega r^2(\tilde{u}_n)\, d\boldsymbol{x} &= \int_\Omega w_n(\boldsymbol{x})\, d\boldsymbol{x}\int_\Omega r^2(u_n(\boldsymbol{x}))\, d\boldsymbol{x}\\
&\leq (1 + \int_\Omega \mathbb{1}_\Omega(\boldsymbol{x})\, d\boldsymbol{x})\int_\Omega r^2(u_n(\boldsymbol{x}))\, d\boldsymbol{x}\\
&= 2\int_\Omega r^2(u_n(\boldsymbol{x}))\, d\boldsymbol{x}.
\end{aligned}$$

This means $\{\tilde{u}_n\}_{n=1}^\infty$ is also a minimizing sequence of equation 17, and it yields

$$\begin{aligned}
d_{W^M}(\tilde{\nu}_n, \mu) &= \sup_{\substack{\|p\|_{\text{Lip}}\leq 1\\ 0\leq p\leq M}} \Big[\int_\Omega p(\boldsymbol{x})\, d\tilde{\nu}_n(\boldsymbol{x}) - \int_\Omega p(\boldsymbol{x})\, d\boldsymbol{x}\Big]\\
&= \sup_{\substack{\|p\|_{\text{Lip}}\leq 1\\ 0\leq p\leq M}} \int_\Omega p(\boldsymbol{x})\Big[\frac{r^2(\tilde{u}_n)(\boldsymbol{x})}{\int_\Omega r^2(\tilde{u}_n(\boldsymbol{x}))\, d\boldsymbol{x}} - \mathbb{1}_\Omega(\boldsymbol{x})\Big]\, d\boldsymbol{x}\\
&= \sup_{\substack{\|p\|_{\text{Lip}}\leq 1\\ 0\leq p\leq M}} \int_\Omega p(\boldsymbol{x})\big[w_n(\boldsymbol{x}) - \mathbb{1}_\Omega(\boldsymbol{x})\big]\, d\boldsymbol{x}\\
&\leq \frac{M}{n}.
\end{aligned}$$

So we get from equation 18 that

$$0 \leq \lim_{n\to\infty} \sup_{\substack{\|p\|_{\text{Lip}}\leq 1\\ 0\leq p\leq M}} \mathcal{J}(\tilde{u}_n, p) \leq \lim_{n\to\infty} 4M\int_\Omega r^2(u_n)\, d\boldsymbol{x} = 0,$$

which means that $\{\tilde{u}_n\}_{n=1}^\infty$ is also a minimizing sequence of equation 5, that is,

$$\lim_{n\to\infty} \mathcal{J}(\tilde{u}_n, p_n) = 0., \tag{15}$$

for some sequence of functions $\{p_n\}_{n=1}^\infty \subset V$. Meanwhile, we have the following properties of $\tilde{u}_n$:

1. The residual sequence $\{r(\tilde{u}_n)\}_{n=1}^{\infty}$ converges to 0 in $L^2(d\mu)$, since

$$\int_\Omega r^2(\tilde{u}_n)\,d\boldsymbol{x} \le 2\int_\Omega r^2(u_n)\,d\boldsymbol{x} \le \frac{2}{n} \to 0, \quad \text{as } n \to \infty$$

2. The renormalized squared residual distributions

$$d\tilde{\nu}_n \triangleq \frac{r^2(\tilde{u}_n)}{\int_\Omega r^2(\tilde{u}_n(\boldsymbol{x}))\,d\boldsymbol{x}}\,d\mu$$

converges to the uniform distribution μ in the Wasserstein distance d_{W^M}.

□

A.3 Replacement of the boundedness condition in Theorem 4

For the boundedness constraint for "test function" p in 4, we prove that it can be removed in our circumstance. And with the following lemma and its following remark, and Theorem 4, we can obtain our main Theorem 1, which is stated again with its assumption in the following.

Assumption. The operator r in equation 7 is a surjection from a function space $E_1(\mathbb{R}^D)$ to $C_c^\infty(\Omega)$, the class of C^∞ functions that are compactly supported on Ω.

Theorem. *Let μ be the Lebesgue measure on $\mathbb{R}^D$, which represents the uniform probability distribution on Ω. In addition, we assume Assumption A1 holds. Then the optimal value of the min-max problem equation 7 is 0. Moreover, there is a sequence $\{u_n\}_{n=1}^{\infty}$ of functions with $r(u_n) \neq 0$ for all n, such that it is an optimization sequence of equation 7, namely,*

$$\lim_{n\to\infty} \mathcal{J}(u_n, p_n) = 0,$$

for some sequence of functions $\{p_n\}_{n=1}^{\infty}$ satisfying the constraints in equation 7. Meanwhile, this optimization sequence has the following two properties:

1. *The residual sequence $\{r(u_n)\}_{n=1}^{\infty}$ of $\{u_n\}_{n=1}^{\infty}$ converges to 0 in $L^2(d\mu)$.*
2. *The renormalized squared residual distributions*

$$d\nu_n \triangleq \frac{r^2(u_n)}{\int_\Omega r^2(u_n(\boldsymbol{x}))\,d\boldsymbol{x}}\,d\mu(\boldsymbol{x})$$

converge to the uniform distribution μ in the Wasserstein distance d_{W^M}.

Although the residue r^2 is renormalized to a probability distribution for the analysis of the algorithm, itself is not a probability distribution, and not treated as so. Actually, in the implementation of our algorithm, the "test function" p is seen as sampling distribution density and the residue r^2 is just the PDE operator (or any kind of objective function whose minimum is 0). In the implementation, we establish p as a generative model, that is, an invertible transform between an unknown distribution (an adversarial distribution to the residual distribution if we think the algorithm as a similarity to GANs) and an "easy-to-sample" distribution such as normal or uniform distribution. So we assume p to be the density function of a probability distribution. Under this assumption, we have the following result.

Lemma 5. *Let Ω be a compact subset of $\mathbb{R}^D$. If a positive function $f : \Omega \to \mathbb{R}$ is K-Lipschitz continuous, and f is the density function of a probability distribution, namely, $\int_\Omega f\,d\boldsymbol{x} = 1$, then there is some constant $M = M(\Omega, K)$, so that $f \le M$. In other words,*

$$f \le M, \quad \forall f \in \mathcal{S} = \big\{f \ge 0 \big| \|f\|_{Lip} \le K, \text{ and } \int_\Omega f\,d\boldsymbol{x} = 1\big\}.$$

Proof. For any $x, y \in \Omega$, we have

$$0 \le f(\boldsymbol{x}) = f(\boldsymbol{x}) - f(\boldsymbol{y}) + f(\boldsymbol{y}) \le K|\boldsymbol{x} - \boldsymbol{y}| + f(\boldsymbol{y}) \le K\mathcal{D}(\Omega) + f(\boldsymbol{y}),$$

where $\mathcal{D}(\Omega)$ is the diameter of Ω. Taking integral with respect to $\boldsymbol{y}$ over Ω on both sides, we have

$$0 \leq f(\boldsymbol{x})\mu(\Omega) \leq K\mathcal{D}(\Omega)\mu(\Omega) + 1,$$

where $\mu(\Omega)$ is the Lebesgue measure (volume) of Ω, that is,

$$0 \leq f(\boldsymbol{x}) \leq K\mathcal{D}(\Omega) + \frac{1}{\mu(\Omega)}.$$

So we have

$$M = M(\Omega, K) = K\mathcal{D}(\Omega) + \frac{1}{\mu(\Omega)}.$$

□

e The converse of this lemma is also true in the sense that if f is bounded by some constant M, then the integral $\int_\Omega f\, d\boldsymbol{x} \leq M\mu(\Omega)$, and f can be renormalized into a probability density function with constant $M\mu(\Omega)$. And similar to boundedness for the gradient (or Lipschitz constant) discussed in section 3.3, a constant renormalizer will not affect the training procedure.

A.4 Deviation of Equation equation 9 and it solution

For a given $r(\boldsymbol{x};\boldsymbol{\theta})$, consider the following minimization problem:

$$\min_{p_{\boldsymbol{\alpha}}>0} \mathcal{L}(p_{\boldsymbol{\alpha}}) = \beta \int_\Omega |\nabla_{\boldsymbol{x}} p_{\boldsymbol{\alpha}}|^2 dx - \int_\Omega r^2(\boldsymbol{x};\boldsymbol{\theta}) p_{\boldsymbol{\alpha}}(\boldsymbol{x}) dx + \lambda \left(\int_\Omega p_{\boldsymbol{\alpha}}(dx) - 1 \right),$$

where the positivity of $p_{\boldsymbol{\alpha}}$ is guaranteed by the KRnet and λ is the Lagrange multiplier for the mass conservation of PDF. Assuming that $\frac{\partial p_{\boldsymbol{\alpha}}}{\partial \boldsymbol{n}} = 0$ on the boundary $\partial\Omega$, where $\boldsymbol{n}$ is a unit normal vector on $\partial\Omega$ pointing outward. We have the first-order variation of $\mathcal{L}(p_{\boldsymbol{\alpha}})$ for a perturbation function $\delta p(\boldsymbol{x})$

$$\begin{aligned}
\delta\mathcal{L} =& 2\beta \int_\Omega \nabla p_{\boldsymbol{\alpha}} \cdot \nabla \delta p d\boldsymbol{x} - \int_\Omega r^2 \delta p d\boldsymbol{x} + \lambda \int_\Omega \delta p d\boldsymbol{x} \\
=& 2\beta \left(\int_{\partial\Omega} \delta p \nabla p_{\boldsymbol{\alpha}} \cdot \boldsymbol{n} d\Gamma - \int_\Omega \delta p \nabla^2 p_{\boldsymbol{\alpha}} d\boldsymbol{x} \right) - \int_\Omega r^2 \delta p d\boldsymbol{x} + \lambda \int_\Omega \delta p(\boldsymbol{x}) d\boldsymbol{x} \\
=& -2\beta \int_\Omega \delta p \nabla^2 p_{\boldsymbol{\alpha}} d\boldsymbol{x} - \int_\Omega r^2 \delta p d\boldsymbol{x} + \lambda \int_\Omega \delta p(\boldsymbol{x}) d\boldsymbol{x} \\
=& -\int_\Omega (2\beta \nabla^2 p_{\boldsymbol{\alpha}} + r^2 - \lambda) \delta p d\boldsymbol{x},
\end{aligned}$$

where we applied integration by parts and the homogeneous Neuman boundary conditions. The optimality condition $\frac{\delta \mathcal{L}}{\delta p} = 0$ yields

$$\begin{cases} 2\beta \nabla^2 p_{\boldsymbol{\alpha}}(\boldsymbol{x}) + r^2(\boldsymbol{x};\boldsymbol{\theta}) - \lambda = 0, & \boldsymbol{x} \in \Omega, \\ \frac{\partial p_{\boldsymbol{\alpha}}}{\partial \boldsymbol{n}} = 0, & \boldsymbol{x} \in \partial\Omega. \end{cases} \tag{19}$$

From the compatibility condition for Neumann problems, we have

$$\int_\Omega (r^2(\boldsymbol{x};\boldsymbol{\theta}) - \lambda) d\boldsymbol{x} = 0, \tag{20}$$

which yields that

$$\lambda = \frac{1}{|\Omega|} \int_\Omega r^2(\boldsymbol{x};\boldsymbol{\theta}) d\boldsymbol{x}.$$

Assume that Ω is a bounded domain with smooth boundary. It can be shown that if $r \in H^k(\Omega)$ and $\partial\Omega \in C^{k+2}$ with $k \in \mathbb{N}$, the solution of equation equation 9 satisfies Taylor (2011)

$$\|p_{\boldsymbol{\alpha}}\|_{H^{k+2}(\Omega)} \leq C \|f\|_{H^k(\Omega)},$$

where $f(\boldsymbol{x}) = (\lambda - r^2)/(2\beta)$ and $C > 0$ is a general constant that does not depend on r. According to the Sobolev Imbedding Theorem Adams & John Fournier (2003),

$$W^{k,1}(\Omega) \to C^{0,1}(\overline{\Omega}),$$

when $D = k - 1$. Thus up to a set of measure zero, we have

$$\|p_{\boldsymbol{\alpha}}\|_{C^{0,1}(\overline{\Omega})} \leq C_1 \|p_{\boldsymbol{\alpha}}\|_{W^{k,1}(\Omega)} \leq C_2 \|p_{\boldsymbol{\alpha}}\|_{H^k(\Omega)},$$

where C_1 and C_2 are general constants independent of $p_{\boldsymbol{\alpha}}$. So $p_{\boldsymbol{\alpha}}$ is Lipschitz continuous when the boundary and $r(\boldsymbol{x})$ are sufficiently smooth. However, this also means that the H_1 regularization used in equation equation 8 induces a weaker constraint than the Lipschitz condition in Lemma 5.

A.5 Supplementary Experiments

About the setting of $s(\boldsymbol{x})$ and $g(\boldsymbol{x})$. The source term $s(\boldsymbol{x})$ is derived by the exact solution, i.e., we can set the source function by plugging the exact solution into the equation to get $s(\boldsymbol{x})$. We set $g(\boldsymbol{x}) = u(\boldsymbol{x})$ since the Dirichlet boundary condition is imposed on $\partial\Omega$.

Parametric Burgers' Equation. We also test the proposed AAS method using parametric PDEs that are commonly used in the design of engineering systems and uncertainty quantification. Specifically, we consider the following parametric Burgers' equation, which is a benchmark problem studied in DeepXDE.

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} = \nu[(\frac{\partial u}{\partial x})^2 + (\frac{\partial u}{\partial y})^2]$$
$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} = \nu[(\frac{\partial v}{\partial x})^2 + (\frac{\partial v}{\partial y})^2]$$
$$x, y \in [0, 1], \text{and } t \in [0, 1]$$

where u and v are the velocities along x and y directions respectively, and $\nu \in (0, 1]$ is a parameter that represents the kinematic viscosity of fluid. Here, the Dirichlet boundary conditions are imposed on all boundaries. The exact solution is obtained as follows.

$$u(x, y, t) = \frac{3}{4} - \frac{1}{4[1 + \exp((-4x + 4y - t)/(32\nu))]},$$
$$v(x, y, t) = \frac{3}{4} + \frac{1}{4[1 + \exp((-4x + 4y - t)/(32\nu))]},$$

The problem setup space is $\boldsymbol{x} = [t, x, y, \nu]$, i.e., $D = 4$. When ν is small, solving this problem is quite challenging. We use the proposed AAS method to train a neural network $u_{\boldsymbol{\theta}}(\boldsymbol{x})$ to approximate the solution over the entire space $\boldsymbol{x} = [t, x, y, \nu] \in [0, 1]^4$. Figure 6 shows the numerical results, which demonstrate that the proposed AAS method is able to accurately solve this parametric Burgers' equation. We can train the models using the strategy as discussed in Remark 2, i.e., we gradually add the data points to the current training set. AAS with fixed $\beta = 5$ means that we use a similar training strategy as DAS-G presented in Tang et al. (2023) with a fixed β, while AAS with decay $\beta = 5$ means that β has a decay scheme at every 100 stages with decay rate 0.9. Adding the data points gradually to the current set of random samples is more stable than that of replacing all data points.

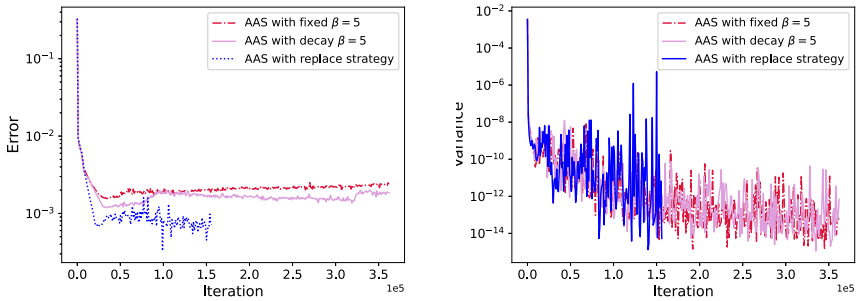

Figure 6: The results of the parametric Burgers' equation. Left: The error behavior. Right: The evolution of the variance.