

Gradient descent globally solves average-case non-resonant physical design problems

Rahul Trivedi*

Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Str. 1, 85748 Garching, Germany.

(Dated: November 5, 2021)

Optimization problems occurring in a wide variety of physical design problems, including but not limited to optical engineering, quantum control, structural engineering, involve minimization of a simple cost function of the state of the system (e.g. the optical fields, the quantum state) while being constrained by the physics of the system. The physics constraints often makes such problems non-convex and thus only locally solvable, leaving open the question of finding the globally optimal design. In this paper, I consider design problems whose physics is described by bi-affine equality constraints, and show that under assumptions on the stability of these constraints and the physical system being non-resonant, gradient descent globally solves a typical physical design problem. The key technical contributions of this paper are (i) outline a criteria that ensure the convergence of gradient descent to an approximate global optima in the limit of large problem sizes, and (ii) use random matrix theory to outline ensembles of physically motivated problems which, on an average, satisfy this convergence criteria.

I. INTRODUCTION

Optimization-based design methodologies have been successfully adapted and applied to solving a wide variety of physical design problems, including but not limited to photonics design [1–6], device-level designs for quantum technologies [7–11], and structural engineering [12–14]. The physics of these system are mathematically captured by bi-affine equality constraints, i.e. constraints of the form $h(x, \theta) = 0$ where h is affine in x for fixed θ and affine in θ for fixed x , relating the system state (x) to the design parameters (θ). The design problem is captured by a cost function of the system state x , which is to be minimized with respect to the parameters θ while enforcing these constraints. In practice, the bi-affine equality constraint allow for an efficient evaluation of the gradient of the cost function with respect to the design parameters [15–19], and thus local optimization algorithms such as gradient descent [20] of quasi-Newton methods [21] can be used to solve them locally. However, the bi-affine constraints also makes the optimization problem non-convex, and hence hard to solve globally. While the locally optimized result suffices for many applications, it leaves open the question of how much more improvement in the device performance can be gained by searching for the global optima.

One prominent approach to answering this question has been to provide lower bounds on the cost function that captures the device performance and is being optimized — while the optimization problem itself is non-convex and hard to solve globally, such bounds can often be computed efficiently by solving a convex problem. Several approaches to setting up this convex problem and calculating these lower bounds using physically motivated convex relaxations [22–26] or by an application of Lagrange duality [27–30] have been recently pursued. In many problems of interest, these lower bounds, computed numerically, are reasonably close to the locally optimized results and thus indicate that the design is near globally-optimal [31, 32]. However, there is no general theoretical

guarantee for the lower bounds obtained by these methods to be tight (i.e. close to the global optima), and consequently they do not always allow for a quantitative assessment of the optimality of the locally optimized design.

In this paper, I pursue a different approach for assessing the optimality of local optimization algorithms — I theoretically analyze the convergence of the gradient descent algorithm when applied to a physical design problem. The results indicate that gradient descent efficiently solves a typical physical design problem i.e. it comes within a specified accuracy of the global optima in a number of steps that only scale polynomially with the problem size. There are two parts to the main result — first, I provide a set of conditions on the physical design problem which guarantee this convergence result. Next, I analyze physically motivated distributions of bi-affine design problems and show that a member of this ensemble, on average, satisfies these conditions and thus is expected to be efficiently solvable by gradient descent. While this work is, to the best of my knowledge, the first rigorous study of convergence of gradient descent for physical design problems, it shares techniques with similar analysis of local optimization algorithms when applied to non-convex problems arising in training of deep neural networks [33, 34], neural-tangent kernels [35] and dynamical models of time-series data [36].

II. NOTATION

For $v \in \mathbb{R}^n$, I will denote by $\|v\|_p$ its l^p norm, and for simplicity use $\|v\|$ for its l^2 norm. For a matrix $M \in \mathbb{R}^{n \times m}$, I will denote by $\|M\|_p$ the induced l^p norm i.e. $\|M\|_p = \sup_{v \in \mathbb{R}^m \setminus \{0\}} \|Mv\|_p / \|v\|_p$. I will denote by I_n the $n \times n$ identity matrix. I will denote by $\|M\|_{\max}$ the maximum magnitude element of M i.e. $\|M\|_{\max} = \max_{i \in [n], j \in [m]} |M_{i,j}|$. I will often write the vector, 0^n (i.e. vector with all elements 0), simply as 0, and the dimensionality of the vector will be evident from the context.

For two vectors $u, v \in \mathbb{R}^n$, $u \odot v \in \mathbb{R}^n$ is their elementwise product defined by $(u \odot v)_i := u_i v_i \forall i \in [n]$. Given a function $f : \text{dom}(f) \subset \mathbb{R} \rightarrow \mathbb{R}$ and a vector $v \in \text{dom}(f)^n$, $f(v)$ will denote a vector obtained on applying f entry-wise

* rahul.trivedi@mpq.mpg.de

to f .

I will use the computer science notation for asymptotic behaviour of sequences. Given a sequence $\{a_n \in \mathbb{R}^+ : n \in \mathbb{N}\}$, $a_n \leq O(f(n))$ for some $f : \mathbb{N} \rightarrow \mathbb{R}^+$ if $\exists c > 0$ such that $a_n \leq cf(n)$ as $n \rightarrow \infty$. $a_n < o(f(n))$ for some $f : \mathbb{N} \rightarrow \mathbb{R}^+$ if $\forall c > 0$, $a_n < cf(n)$ as $n \rightarrow \infty$. In particular, $n^{o(1)}$ will be used to denote a function whose growth is slower than n^α for any $\alpha > 0$. $a_n \geq \Omega(f(n))$ for some $f : \mathbb{N} \rightarrow \mathbb{R}^+$ if $\exists c > 0$ such that $a_n \geq cf(n)$ as $n \rightarrow \infty$. $a_n = \Theta(f(n))$ if $a_n \leq O(f(n))$ and $a_n \geq \Omega(f(n))$.

III. SUMMARY OF RESULTS

I first introduce a definition of a physical design problem, along with some terminology that I use throughout this paper. I introduce an abstract definition of a *physical system*, and a *physical design problem*.

Definition 1 A physical system with state size n and parameter size m is a map $\varphi : \text{dom}(\varphi) \rightarrow \mathbb{R}^n$ specified by a tuple (A, B, b) where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $b \in \mathbb{R}^n$ and

$$\text{dom}(\varphi) = \{\theta \in \mathbb{R}^m \mid A + \text{diag}(B\theta) \text{ is invertible}\},$$

and $\forall \theta \in \mathbb{R}^m$,

$$\varphi(\theta) = (A + \text{diag}(B\theta))^{-1}b.$$

A is the physics matrix, B is the selection matrix and b is the source vector corresponding to the physical system.

Definition 2 Given a physical system $\varphi \equiv (A, B, b)$ of state size n , and a vector $c \in \mathbb{R}^n$, the adjoint system is a map $\text{ad}[\varphi] : \text{dom}(\varphi) \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that $\forall \theta \in \mathbb{R}^m$,

$$\text{ad}[\varphi](\theta, c) = (A + \text{diag}(B\theta))^{-T}c.$$

Definition 3 A physical design problem of state size n and parameter size m is specified by a tuple (f, c, φ) where $f : \mathbb{R} \rightarrow \mathbb{R}$ is the cost function, $c \in \mathbb{R}^n$ is the overlap vector, φ is a physical system of state size n and parameter size m and it corresponds to solving the following constrained optimization problem:

$$\underset{\theta \in \text{dom}(\varphi)}{\text{minimize}} \quad f(c^T \varphi(\theta)),$$

To solve this problem using local optimization algorithm, it is essential to be able to efficiently compute the gradient of the cost function. It is well known that this can be done using the the maps corresponding to the physical system and the adjoint system [15–19]. This is made explicit in the following lemma, which can be straightforwardly proved using the chain rule.

Lemma 1 For a physical design problem (f, c, φ) , then for the map $f(c^T \varphi(\cdot)) : \text{dom}(\varphi) \rightarrow \mathbb{R}$, the gradient at $\theta \in \text{dom}(\varphi)$ is given by

$$\nabla_\theta f(c^T \varphi(\theta)) = -f'(c^T \varphi(\theta))B_\varphi^T(\varphi(\theta) \odot \text{ad}[\varphi](\theta, c)).$$

Algorithm 1 Gradient descent for solving a physical design problem

Input: A physical design problem (f, c, φ) , an initial set of design parameters $\theta_0 \in \mathbb{R}^m$, a gradient descent step size $\eta \in (0, \infty)$ and number of gradient descent steps $T \in \mathbb{N}$.

Output: The optimized design parameters $\theta^* \in \mathbb{R}^m$.

Initialisation :

If $\theta_0 \notin \text{dom}(\varphi)$, then declare FAIL, else set $x_0 = \varphi(\theta_0) = (A_\varphi + \text{diag}(B_\varphi \theta_0))^{-1}b_\varphi$.

LOOP Process

for $t = 1$ to $T - 1$ **do**

 Compute the adjoint state $a_{t-1} = \varphi_c^{\text{adj}}(\theta_{t-1})$.

 Compute the gradient $g_{t-1} = -f'(c^T x_{t-1})(\varphi_c^{\text{adj}}(\theta_{t-1}) \odot \varphi(\theta_{t-1}))$

 Perform the gradient descent step $\theta_t := \theta_{t-1} - \eta g_{t-1}$.

if $(\theta_t \notin \text{dom}(\varphi))$ **then**

 Declare FAIL.

end if

end for

return θ_T

Throughout this paper, I will be interested in performing an analysis of gradient descent (explicitly described in algorithm 1) not for one specific problem instance, but for problem instance with large state sizes. More formally, I will consider a family of physical design problems with larger and larger state sizes and assess how gradient descent performs on a specific instance of this family.

Definition 4 A family of physical design problems is a sequence $\{(f, c_n, \varphi_n) : n \in \mathbb{N}\}$ of physical design problems, where (f, c_n, φ_n) is a physical design problem of state size n and the cost function $f : \mathbb{R} \rightarrow \mathbb{R}$ is independent of n .

I next introduce a set of conditions on a given family of physical design problems, which will be central to the analysis of the convergence of gradient descent. These conditions are expressed in terms of the scalings of the spectrum of the physics matrix with the state size of the physical system, as well as on the norms of the source and overlap vectors. The definition I provide below makes concrete the physical expectation that the norm of the physics matrix of systems encountered in practice only grows polynomially with the state size of the system, and typically the source and overlap vectors only affect a constant number of state variables and hence have norms upper bounded by a constant. I also assume a condition on the initial gradient of the cost function — since in practice the initial design is picked randomly and is thus not locally optimal, we assume that the initial gradient of the cost function is large (i.e. lower bounded by a function that grows polynomially with the state size).

Definition 5 A family of physical design problems $\{(f, c_n, \varphi_n \equiv (A_n, B_n, b_n)) : n \in \mathbb{N}\}$ is said to satisfy the (α, γ) -asymptotic convergence conditions if

(a) f is L -Lipschitz smooth¹, μ -strongly convex² and is

¹ A differentiable function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ is L -Lipschitz smooth if $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

² A function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ is μ -strongly convex if $f(y) \geq f(x) +$

bounded from below i.e. $f^* = \min_{x \in \mathbb{R}} f(x) > -\infty$.

- (b) $\|A_n^{-1}\|_2 \leq O(n^\gamma)$,
- (c) $\|B_n\|_\infty, \|B_n\|_1, \|B_n\|_2 \leq O(1)$,
- (d) $\|b_n\|_\infty, \|b_n\|_1 \leq O(1)$,
- (e) $\|c_n\|_\infty, \|c_n\|_1 \leq O(1)$,
- (f) $\|B_n^T(A_n^{-1}b_n \odot A_n^{-T}c_n)\|_2^2 \geq \Omega(n^{\alpha+\gamma})$ and
- (g) $|c_n^T A_n^{-1}b_n| \leq O(1)$.

Typical physical design problems are expected to be non-resonant i.e. the state vector is expected to not become very large during the optimization trajectory. It is empirically observed in many settings that local optimization algorithms perform very well when applied to such non-resonant design problems, while resonant design problems are usually much harder to solve. This consideration has to be accounted for in the analysis of gradient descent — the next definition that I provide makes the intuition behind a non-resonant design mathematically precise, and will be another important ingredient in the analysis of the optimality of gradient descent.

Definition 6 (Non-resonant design parameters)

Given a family of physical design problems $\mathcal{F} = \{(f, c_n, \varphi_n) : n \in \mathbb{N}\}$, a sequence of design parameters $\{\theta_n \in \text{dom}(\varphi_n) : n \in \mathbb{N}\}$ is said to be non-resonant with respect to the family \mathcal{F} if $\|\varphi_n(\theta_n)\|_\infty \leq O(n^{o(1)})$ and $\|\text{adj}[\varphi_n](\theta_n, c_n)\|_\infty \leq O(n^{o(1)})$.

I now present the first result of this paper — a family of physical design problems is efficiently solvable by gradient descent, under the assumption that the design parameters generated during the algorithm are non-resonant, if it satisfies the asymptotic convergence conditions.

Theorem 1 Let $\mathcal{F} := \{(f, c_n, \varphi_n) : n \in \mathbb{N}\}$ be a family of physical design problems that satisfies the α, γ -asymptotic convergence conditions (definition 5) with $\alpha > 1/4, \gamma < 3\alpha$, and for $n \in \mathbb{N}$, let $\Theta_n = \{\theta_n^1, \theta_n^2 \dots \theta_n^{T_n}\}$ be the design parameters generated by gradient descent (algorithm 1) under the assumption that it does not fail when applied on (f, c_n, φ_n) with step size $\eta_n \leq O(n^{-3\gamma+\alpha-1})$ for T_n steps starting with initial design of $\theta_n^0 = 0$. If all the sequences $\{\theta_n \in \Theta_n : n \in \mathbb{N}\}$ are non-resonant with respect to \mathcal{F} , then $f(c_n^T \varphi_n(\theta_n^{T_n})) - f^* \leq \varepsilon$ for T_n chosen such that $T_n = \Theta(n^{1-2(\alpha-\gamma)} \log(\varepsilon^{-1}))$.

In the next two propositions, I provide families of design problems which are constructed by choosing the physics matrices randomly from a distribution. By analyzing an average-case problem picked from these distributions, I show that it satisfies the asymptotic convergence conditions. The first family of problem is one in which the inverse of the physics matrix is a matrix from the random gaussian ensemble, and this

example is inspired from wave design problems [1–4], where often the initial design is a random scattering media whose properties are known to be captured by random matrices [37].

Problem 1 Given a cost function $f : \mathbb{R} \rightarrow \mathbb{R}$ being an L -Lipschitz smooth and μ -strongly convex function that is bounded from below, define a family of physical design problems $\mathcal{F} := \{(f, c_n, \varphi_n \equiv (A_n := G_n^{-1}, B_n := I_n, b_n)) : n \in \mathbb{N}\}$ where

- $\|b_n\|_\infty, \|b_n\|_1, \|c_n\|_\infty, \|c_n\|_1 = \Theta(1)$ and,
- $G_n \in \mathbb{R}^{n \times n}$ is a matrix where each entry is independently drawn from the standard normal distribution.

Proposition 1 For the problem of state size n , $(f, c_n, \varphi_n \equiv (A_n, B_n, b_n))$, picked from the family of problems defined in problem 1, it is true that

- (a) With probability $1 - 2 \exp(-\varepsilon^2/2)$, $\|A_n^{-1}\|_2 \leq 2\sqrt{n} + \varepsilon$,
- (b) $E(\|B_n^T(A_n^{-1}b_n \odot A_n^{-T}c_n)\|_2^2) \geq \Omega(n)$,
- (c) $E(|c_n^T A_n^{-1}b_n|^2) \leq O(1)$,

and consequently, this family of problems on an average satisfies the $(1/2, 1/2)$ -asymptotic convergence conditions.

Proposition 1 thus shows that theorem 1 is applicable on average for a design problem drawn from the distribution of problems defined in problem 1, and thus guarantees an average case convergence of gradient descent.

The second family of problems is one in which I fix the scaling of the singular values of the physics matrix with the state size, and generate the left and right singular vectors randomly — the physical motivation behind this construction is that the scaling of the spectrum of the physics matrix with the problem size is often fixed by the derivative operators and boundary conditions appearing in the physical laws, while the precise singular vectors depend on the details of the (initial and randomly chosen) design parameters. Under some assumptions on the scalings of the singular values with the problem size, the asymptotic convergence conditions (definition 5) are shown to be satisfied on an average.

Problem 2 Given a cost function $f : \mathbb{R} \rightarrow \mathbb{R}$ being an L -Lipschitz smooth and μ -strongly convex function that is bounded from below and $\gamma \geq 1/2$, define a family of physical design problems $\{(f, c_n, \varphi_n \equiv (A_n := R_n \text{diag}(s_n) Q_n^T, B_n := I_n, b_n)) : n \in \mathbb{N}\}$ where

- $\|b_n\|_\infty, \|b_n\|_1, \|c_n\|_\infty, \|c_n\|_1 = \Theta(1)$,
- $s_n \in (0, \infty)^n$ with $\|1/s_n\|_\infty \leq O(n^\gamma)$ and $\|1/s_n\| = \Theta(n)$,
- $R_n, Q_n \in \mathbb{R}^{n \times n}$ are drawn uniformly at random from the Haar measure over orthogonal matrices.

Proposition 2 For the problem of state size n , $(f, c_n, \varphi_n \equiv (A_n, B_n, b_n))$, picked from the family of problems defined in problem 2,

$$\nabla f(x)^T(y - x) + \sigma \|x - y\|^2/2.$$

- (a) $\|A_n^{-1}\|_2 \leq O(n^\gamma)$,
- (b) $E(\|B_n^T(A_n^{-1}b_n \odot A_n^{-T}c_n)\|_2^2) \geq \Omega(n)$,
- (c) $E(|c_n^T A_n^{-1} b_n|^2) \leq O(1)$,

and consequently, this family of problems on an average satisfies the $1 - \gamma, \gamma$ -asymptotic convergence conditions.

Consequently using theorem 1, it follows that for $\gamma \in [1/2, 3/4)$, a design problem drawn from the distribution of design problems defined in problem 1 on an average is efficiently solvable by gradient descent.

Finally, I show that under some further assumptions on the physics matrix of the physical system, it can be shown that all the designs generated during the gradient descent algorithm are non-resonant, and consequently gradient descent can be shown to converge to the global optima without the any additional assumptions on the gradient descent trajectory.

Theorem 2 Let $\mathcal{F} := \{(f, c_n, \varphi_n \equiv (A_n, B_n, b_n)) : n \in \mathbb{N}\}$ be a family of physical design problems that satisfies the α, γ -asymptotic convergence conditions (definition 5) with $\alpha > 1/2, \gamma < 3\alpha$, and also satisfies $\|A_n^{-1}\|_{\max} \leq O(n^{o(1)})$, then gradient descent when applied on (f, c_n, φ_n) , if it does not fail, produces a design θ_n^* such that $f(c^T \varphi_n(\theta_n^*)) - f^* \leq \varepsilon$ in $T = \Theta(n^{1-2(\alpha-\gamma)} \log(\varepsilon^{-1}))$ steps.

The remainder of this paper is devoted to detailed proofs of these statements — the proof of theorem 1 is provided in section IV A, section IV B is dedicated to the proofs of propositions 1 and 2 and the proof of theorem 2 is provided in section IV C.

IV. DETAILED PROOFS

A. Convergence of gradient descent

I begin by establishing an asymptotic property concerning the stability of a physical system — the object of interest here is to study how a physical system behaves when the design parameters are perturbed slightly.

Lemma 2 Let $\{\varphi_n \equiv (A_n, B_n, b_n) : n \in \mathbb{N}\}$ be a sequence of physical systems which satisfies $\|B_n\|_\infty \leq O(1)$ and $\|A^{-1}\|_2 \leq O(n^\gamma)$ for some $\gamma > 0$, then for all sequences $\{\theta_n \in \text{dom}(\varphi_n) : n \in \mathbb{N}\}$ such that $\|\theta_n\|_\infty \leq O(n^{o(1)-\alpha-\gamma})$ for some $\alpha > 0$, $\|(A_n + \text{diag}(B_n \theta_n))^{-1}\|_2 \leq O(n^\gamma)$.

Proof: It follows straightforwardly that for $n \in \mathbb{N}$,

$$(A_n + \text{diag}(B_n \theta_n))^{-1} = A_n^{-1} - A_n^{-1} \text{diag}(B_n \theta_n) (A_n + \text{diag}(B_n \theta_n))^{-1}.$$

From the triangle inequality, I obtain that

$$\|(A_n + \text{diag}(B_n \theta_n))^{-1}\|_2 \leq \|A_n^{-1}\|_2 + \|A_n^{-1}\|_2 \|B_n \theta_n\|_\infty \|(A_n + \text{diag}(B_n \theta_n))^{-1}\|_2.$$

By assumption, $\|A_n^{-1}\|_2 \leq O(n^\gamma)$ for $\gamma > 0$, and $\|B_n \theta_n\|_\infty \leq \|B_n\|_\infty \|\theta_n\|_\infty \leq O(n^{o(1)-\gamma-\alpha})$. Therefore,

$$(1 - O(n^{o(1)-\alpha})) \|(A_n + \text{diag}(B_n \theta_n))^{-1}\|_2 \leq O(n^\gamma),$$

from which the lemma statement follows. \square

For completeness, I provide an additional lemma with some standard properties of Lipschitz continuous and strongly convex functions that I will use in the following proofs.

Lemma 3 Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a twice-differentiable, L -Lipschitz smooth and μ -strongly convex function such that $f^* = \min_{x \in \mathbb{R}} f(x) \geq -\infty$, then

- (a) $\forall x, y \in \mathbb{R}$,

$$f(x) \leq f(y) + \nabla f(y)(x - y) + \frac{L}{2}|x - y|^2.$$

- (b) $\forall x \in \mathbb{R}$,

$$2\mu(f(x) - f^*) \leq |\nabla f(x)|^2.$$

- (c) $\forall x \in \mathbb{R}$,

$$|f'(x)|^2 \leq \frac{2L^2}{\mu}(f(x) - f^*).$$

Proof:

- (a) Since a L -Lipschitz smooth function by definition satisfies $|f'(x) - f'(y)| \leq L|x - y| \forall x, y \in \mathbb{R}$, it follows that $\forall x \in \mathbb{R}$, $|f''(x)| \leq L$. From Taylor's theorem, it follows that

$$f(x) = f(y) + f'(y)(x - y) + \frac{1}{2} \int_y^x f''(s)(x - s) ds.$$

Since $\forall s \in [x, y]$, $f''(s) \leq |f''(s)| \leq L$, it follows that

$$\begin{aligned} f(x) &\leq f(y) + f'(y)(x - y) + \frac{L}{2} \int_y^x (x - s) ds \\ &= f(y) + f'(y)(x - y) + \frac{L}{2}|x - y|^2. \end{aligned}$$

- (b) Since f is μ -strongly convex, $\forall x, y \in \mathbb{R}$

$$f(x) \geq f(y) + f'(y)(x - y) + \frac{\mu}{2}|x - y|^2.$$

and hence $\forall y \in \mathbb{R}$

$$\min_{x \in \mathbb{R}} f(x) \geq \min_{x \in \mathbb{R}} \left(f(y) + f'(y)(x - y) + \frac{\mu}{2}|x - y|^2 \right),$$

from which it follows that $\forall y \in \mathbb{R}$

$$f^* \geq f(y) - \frac{1}{2\mu}|f'(y)|^2.$$

(c) Let $x^* = \operatorname{argmin}_{x \in \mathbb{R}} f^*$. I note that from stationarity conditions, $f'(x^*) = 0$. From the L -Lipschitz continuity, it follows that $|f'(x)|^2 \leq L^2(x - x^*)^2 \forall x \in \mathbb{R}$. Furthermore, from strong convexity of f , it follows that $\forall x \in \mathbb{R}$,

$$f(x) \geq f^* + \frac{\mu}{2}|x - x^*|^2.$$

Therefore, $\forall x \in \mathbb{R}$, $|f'(x)|^2 \leq 2L^2/\mu(f(x) - f^*) \square$.

Next, I analyze the gradient descent algorithm (algorithm 1). The next three lemmas characterize the decrease in the cost function on taking a gradient descent step.

Lemma 4 *Let $(f, c, \varphi \equiv (A, B, b))$ be a physical design problem, and let f be L -Lipschitz smooth, then $\forall \theta \in \operatorname{dom}(\varphi)$ and $\eta > 0$ such that $\theta' := \theta - \eta \nabla_{\theta} f(c^T \varphi(\theta)) \in \operatorname{dom}(\varphi)$,*

$$f(c^T \varphi(\theta')) \leq f(c^T \varphi(\theta)) - \eta (f'(c^T \varphi(\theta))^2 \times (\|B^T v(\theta)\|_2^2 + \varepsilon_1(\theta, \theta') + \eta \varepsilon_2(\theta, \theta'))),$$

where $v(\theta) := \varphi(\theta) \odot \operatorname{ad}[\varphi](\theta, c)$

$$\varepsilon_1(\theta, \theta') := \operatorname{ad}[\varphi](\theta, c)^T \operatorname{diag}(BB^T v(\theta))(\varphi(\theta') - \varphi(\theta)), \quad (1a)$$

$$\varepsilon_2(\theta, \theta') := -\frac{L}{2} (\operatorname{ad}[\varphi](\theta, c)^T \operatorname{diag}(BB^T v(\theta)) \varphi(\theta'))^2. \quad (1b)$$

Proof: Since f is L -Lipschitz smooth, it follows that

$$f(c^T \varphi(\theta')) \leq f(c^T \varphi(\theta)) + f'(c^T \varphi(\theta)) c^T (\varphi(\theta') - \varphi(\theta)) + \frac{L}{2} (c^T (\varphi(\theta) - \varphi(\theta')))^2. \quad (2)$$

Furthermore, since

$$\varphi(\theta') - \varphi(\theta) = -(A + \operatorname{diag}(B\theta))^{-1} \operatorname{diag}(B(\theta' - \theta)) \varphi(\theta'),$$

I obtain that

$$\begin{aligned} & c^T (\varphi(\theta') - \varphi(\theta)) \\ &= -\operatorname{ad}[\varphi](\theta, c)^T \varphi(\theta') \\ &= -\eta f'(c^T \varphi(\theta)) \operatorname{adj}[\varphi](\theta, c)^T \operatorname{diag}(BB^T v(\theta)) \varphi(\theta'). \end{aligned} \quad (3)$$

Expressing $\varphi(\theta') = \varphi(\theta) + (\varphi(\theta') - \varphi(\theta))$, I obtain that

$$\begin{aligned} c^T (\varphi(\theta') - \varphi(\theta)) &= -\eta f'(c^T \varphi(\theta)) (\|B^T v(\theta)\|_2^2 + \\ & \operatorname{adj}[\varphi](\theta, c)^T \operatorname{diag}(BB^T v(\theta)) (\varphi(\theta') - \varphi(\theta))). \end{aligned} \quad (4)$$

Substituting Eqs. 3 and 4 into the Eq. 2, I obtain the lemma statement. \square .

Lemma 5 *Consider a family of physical design problems $\mathcal{F} := \{(f, c_n, \varphi_n \equiv (A_n, B_n, b_n))\}$ satisfying the α, γ -asymptotic convergence conditions (definition 5). Let $\eta > 0$, and let $\{\theta_n \in \operatorname{dom}(\varphi_n) : n \in \mathbb{N}\}$ be a sequence non-resonant with respect to \mathcal{F} with $\|\theta_n\|_{\infty} \leq O(n^{\alpha(1)-\alpha-\gamma})$. Furthermore, suppose that the sequence*

$\{\theta'_n := \theta_n - \eta \nabla_{\theta} f(c_n^T \varphi_n(\theta_n)) : n \in \mathbb{N}\}$ satisfies $\|\theta'_n\|_{\infty} \leq O(n^{\alpha(1)-\alpha-\gamma})$, then $|\varepsilon_1(\theta_n, \theta'_n)| \leq O(n^{\alpha(1)+\gamma-\alpha+1/2})$, where ε_1 is defined in lemma 4.

Proof: From the definition of ε_1 it follows that $\forall n \in \mathbb{N}$,

$$|\varepsilon_1(\theta_n, \theta'_n)| \leq \sqrt{n} \|\operatorname{ad}[\varphi_n](\theta_n, c_n)^T \operatorname{diag}(B_n B_n^T v_n(\theta_n))\|_{\infty} \times \|\varphi_n(\theta'_n) - \varphi_n(\theta_n)\|_2, \quad (5)$$

where $v_n(\theta_n) = \operatorname{ad}[\varphi_n](\theta_n, c_n) \odot \varphi_n(\theta_n)$. Since the sequence $\{\theta_n : n \in \mathbb{N}\}$ is non-resonant, $\|\operatorname{ad}[\varphi_n](\theta_n, c_n)^T \operatorname{diag}(B_n B_n^T v_n(\theta_n))\|_{\infty} \leq O(n^{\alpha(1)})$. Furthermore, $\forall n \in \mathbb{N}$

$$\begin{aligned} & \|\varphi_n(\theta_n) - \varphi_n(\theta'_n)\|_2 \leq \\ & \|(A_n + \operatorname{diag}(B_n \theta_n))^{-1}\|_2 \|B_n(\theta_n - \theta'_n)\|_{\infty} \|\varphi_n(\theta'_n)\|_2 \end{aligned}$$

It follows from lemma 2 that $\|(A_n + \operatorname{diag}(B_n \theta_n))^{-1}\|_2 \leq O(n^{\gamma})$ and $\|\varphi_n(\theta'_n)\|_2 \leq \|(A_n + \operatorname{diag}(B_n \theta_n))^{-1}\|_2 \sqrt{\|b\|_{\infty} \|b\|_1} \leq O(n^{\gamma})$. Finally, $\|B_n(\theta_n - \theta'_n)\|_{\infty} \leq \|B_n\|_{\infty} (\|\theta_n\|_{\infty} + \|\theta'_n\|_{\infty}) \leq O(n^{\alpha(1)-\alpha-\gamma})$ and thus $\|\varphi_n(\theta'_n) - \varphi_n(\theta_n)\|_2 \leq O(n^{\alpha(1)+\gamma-\alpha})$. Using these estimates the lemma statement follows.

Lemma 6 *Consider a family of physical design problems $\mathcal{F} := \{(f, c_n, \varphi_n \equiv (A_n, B_n, b_n))\}$ satisfying the α, γ -asymptotic convergence conditions (definition 5). Let $\eta > 0$ and $\{\theta_n \in \operatorname{dom}(\varphi_n) : n \in \mathbb{N}\}$ be a sequence non-resonant with respect to \mathcal{F} with $\|\theta_n\|_{\infty} \leq O(n^{\alpha(1)-\alpha-\gamma})$. Furthermore, suppose that the sequence $\{\theta'_n := \theta_n - \eta \nabla_{\theta} f(c_n^T \varphi_n(\theta_n)) : n \in \mathbb{N}\}$ satisfies $\|\theta'_n\|_{\infty} \leq O(n^{\alpha(1)-\alpha-\gamma})$, then $|\varepsilon_2(\theta_n, \theta'_n)| \leq O(n^{\alpha(1)+4\gamma})$, where ε_2 is defined in lemma 4.*

Proof: I note that $\forall n \in \mathbb{N}$, $|\varepsilon_2(\theta_n)| \leq L/2 \|\operatorname{adj}[\varphi_n](\theta_n, c_n)\| \|\varphi_n(\theta_n)\| \|B_n B_n^T v_n(\theta_n)\|_{\infty}$, where, as in lemma 5, $v_n(\theta_n) := \operatorname{ad}[\varphi_n](\theta_n, c_n) \odot \varphi_n(\theta_n) \forall n \in \mathbb{N}$. It follows from lemma 2 and the observation that $\|b\| \leq \sqrt{\|b\|_1 \|b\|_{\infty}} \leq O(1)$ that $\|\varphi_n(\theta_n)\| \leq O(n^{\gamma})$. Similarly, $\|\operatorname{ad}[\varphi_n](\theta_n, c_n)\| \leq O(n^{\gamma})$. Furthermore, $\|B_n B_n^T\|_{\infty} \leq \|B_n\|_{\infty} \|B_n\|_1 \leq O(1)$ and since $\{\theta_n : n \in \mathbb{N}\}$ is non-resonant, $\|\operatorname{ad}[\varphi_n](\theta_n, c_n)\|_{\infty} \leq O(n^{\alpha(1)})$. From these estimates, the lemma statement follows.

Lemma 7 *Consider a family of physical design problems $\mathcal{F} := \{(f, c_n, \varphi_n \equiv (A_n, B_n, b_n))\}$ satisfying the α, γ -asymptotic convergence conditions (definition 5) with $3\alpha > \gamma$ and $\|A_n^{-T} c_n\|_{\infty} \leq O(n^{\alpha(1)})$. Let $\{\theta_n \in \operatorname{dom}(\varphi_n) : n \in \mathbb{N}\}$ be a sequence non-resonant with respect to \mathcal{F} such that $\|\theta_n\|_{\infty} \leq O(n^{\alpha(1)-\alpha-\gamma})$, then $\|B_n^T (\varphi_n(\theta_n) \odot \operatorname{ad}[\varphi_n](\theta_n, c_n))\|_2^2 \geq \Omega(n^{\alpha+\gamma})$.*

Proof: For notational convenient, let $v_n(\theta) = \varphi_n(\theta) \odot \operatorname{ad}[\varphi_n](\theta, c_n) \forall \theta \in \operatorname{dom}(\varphi_n)$, $n \in \mathbb{N}$. I first upper bound $\|B_n^T (v_n(\theta_n) - v_n(0))\|$. Notice that

$$\|B_n^T (v_n(\theta_n) - v_n(0))\| \leq \|B_n\|_2 \|v_n(\theta_n) - v_n(0)\|.$$

Furthermore,

$$\begin{aligned} \|v_n(\theta_n) - v_n(0)\| &\leq \\ &\|\varphi_n(\theta_n)\|_\infty \|\text{ad}[\varphi_n](\theta_n, c_n) - \text{ad}[\varphi_n](0, c_n)\| + \\ &\|\text{ad}[\varphi_n](0, c_n)\|_\infty \|\varphi_n(\theta_n) - \varphi_n(0)\|. \end{aligned}$$

Note that since $\{\theta_n : n \in \mathbb{N}\}$ is non-resonant, $\|\varphi_n(\theta_n)\|_\infty \leq O(n^{o(1)})$ and by assumption $\|\text{ad}[\varphi_n](0, c_n)\|_\infty = \|A_n^{-T}c_n\|_\infty \leq O(n^{o(1)})$. Furthermore, $\forall n \in \mathbb{N}$

$$\|\varphi_n(\theta_n) - \varphi_n(0)\| \leq \|A_n^{-1}\|_2 \|B_n \theta_n\|_\infty \|\varphi_n(\theta_n)\|.$$

I note that by assumption $\|A_n^{-1}\|_2 \leq O(n^\gamma)$, $\|B_n \theta_n\| \leq O(n^{o(1)-\alpha-\gamma})$ and $\|\varphi_n(\theta_n)\| \leq \|(A_n + \text{diag}(B_n \theta_n))^{-1}\|_2 \|b_n\| \leq O(n^\gamma)$. Consequently, $\|\varphi_n(\theta_n) - \varphi_n(0)\| \leq O(n^{o(1)+\gamma-\alpha})$. Similarly, $\|\text{ad}[\varphi_n](\theta_n, c_n) - \text{ad}[\varphi_n](0, c_n)\| \leq O(n^{o(1)+\gamma-\alpha})$, which yields

$$\|B_n^T(v_n(\theta_n) - v_n(0))\| \leq O(n^{o(1)+\gamma-\alpha}).$$

Finally, I note that from the triangle inequality

$$\|B_n^T v_n(\theta_n)\| \geq \|B_n^T v_n(0)\| - \|B_n^T(v_n(\theta_n) - v_n(0))\|.$$

Since, by assumption, $\|B_n^T v_n(0)\| \geq \Omega(n^{(\alpha+\gamma)/2})$ and $\gamma < 3\alpha$, I obtain that $\|B_n^T v_n(\theta_n)\| \geq \Omega(n^{(\alpha+\gamma)/2})$, from which the lemma statement follows. \square .

Theorem 1 (Restated) *Let $\mathcal{F} := \{(f, c_n, \varphi_n) : n \in \mathbb{N}\}$ be a family of physical design problems that satisfies the α, γ -asymptotic convergence conditions (definition 5) with $\alpha > 1/4, \gamma < 3\alpha$, and for $n \in \mathbb{N}$, let $\Theta_n = \{\theta_n^1, \theta_n^2 \dots \theta_n^{T_n}\}$ be the design parameters generated by gradient descent (algorithm 1) under the assumption that it does not fail when applied on (f, c_n, φ_n) with step size $\eta_n \leq O(n^{-3\gamma+\alpha-1})$ for T_n steps starting with initial design of $\theta_n^0 = 0$. If all the sequences $\{\theta_n \in \Theta_n : n \in \mathbb{N}\}$ are non-resonant with respect to \mathcal{F} , then $f(c_n^T \varphi_n(\theta_n^{T_n})) - f^* \leq \varepsilon$ for T chosen such that $T_n = \Theta(n^{1-2(\alpha-\gamma)} \log(\varepsilon^{-1}))$.*

Proof: Consider the trajectory $\{\theta_n^1 \dots \theta_n^{T_n}\}$ generated by gradient descent when applied for T_n steps on the problem (f, c_n, φ_n) starting from $\theta_n^0 = 0$ and with step size η_n . I will first analyze the final cost function achieved under the assumption that $\forall t \in [T_n], \|\theta_n^t\|_\infty \leq O(n^{o(1)-\alpha-\gamma})$, and show that it can be made ε -close to f^* after $T_n = \Theta(n^{1-2(\alpha-\gamma)} \log(\varepsilon^{-1}))$ gradient descent steps. Then, I will show that this assumption is valid for all gradient descent steps.

I note with this assumption and the assumption that gradient descent algorithm generates a trajectory that is non-resonant, lemmas 4-7 are applicable. From lemmas 5 and 7, it is easy to see that if $\alpha > 1/4$, then $\forall t \in \{0, 1, 2 \dots T_n - 1\}$,

$$\|B_n^T(\varphi_n(\theta_n^t) \cdot \text{ad}[\varphi_n](\theta_n^t, c_n))\|^2 + 2\varepsilon_1(\theta_n^t, \theta_n^{t+1}) \geq \Omega(n^{\alpha+\gamma}).$$

Similarly, choosing $\eta_n \leq O(n^{-3\gamma+\alpha-1})$, it follows from lem-

mas 6 and 7 that $\forall t \in \{0, 1, 2 \dots T_n - 1\}$

$$\|B_n^T(\varphi_n(\theta_n^t) \cdot \text{ad}[\varphi_n](\theta_n^t, c_n))\|^2 + 2\varepsilon_2(\theta_n^t, \theta_n^{t+1}) \geq \Omega(n^{\alpha+\gamma}).$$

Consequently, from lemma 4, I obtain that $\forall t \in [T_n]$

$$f(c_n^T \varphi_n(\theta_n^t)) \leq f(c_n^T \varphi_n(\theta_n^{t-1})) - \eta_n f'(c_n^T \varphi_n(\theta_n^{t-1}))^2 \Omega(n^{\alpha+\gamma}),$$

From lemma 3b it follows that

$$f(c_n^T \varphi_n(\theta_n^t)) - f^* \leq (f(c_n^T \varphi_n(\theta_n^{t-1})) - f^*)(1 - \eta_n \Omega(n^{\alpha+\gamma})).$$

where $f^* = \min_{x \in \mathbb{R}} f(x)$. Noting that, from lemma 3b along with asymptotic convergence condition (g), it follows that

$$\begin{aligned} f(c_n^T \varphi_n(\theta_n^0)) - f^* &= f(c_n^T A_n^{-1} b_n) - f^* \\ &\leq \frac{L}{2} (c_n^T A_n^{-1} b_n - x^*)^2 \\ &\leq O(1), \end{aligned}$$

where $x^* = \text{argmin}_{x \in \mathbb{R}} f(x)$ and therefore

$$f(c_n^T \varphi_n(\theta_n^{T_n})) - f^* \leq O(1)(1 - \eta_n \Omega(n^{\alpha+\gamma}))^{T_n}.$$

Consequently, using $T_n = \Theta(n^{1-2(\alpha-\gamma)} \log(\varepsilon^{-1}))$, I obtain that $f(c_n^T \varphi_n(\theta_n^{T_n})) - f^* \leq \varepsilon$.

Next, I verify that the assumption $\|\theta_n^t\|_\infty \leq O(n^{o(1)-\alpha-\gamma})$ holds for all $t \in [T_n]$. I notice that for all $t \in [T_n]$, the norm of θ_n^t can be upper bounded by the sum of norms of gradients in the previous steps, multiplied by the step size. Using lemma 1

$$\begin{aligned} \|\theta_n^t\|_\infty &\leq \\ &\eta_n \sum_{t'=0}^{t-1} |f'(c_n \varphi_n(\theta_n^{t'}))| \|B_n^T(\varphi_n(\theta_n^{t'}) \odot \text{ad}[\varphi_n](\theta_n^{t'}, c_n))\|_\infty. \end{aligned}$$

Furthermore, using the fact that $\|B_n\|_1 \leq O(1)$, and that the gradient descent trajectory is assumed to be non-resonant, $\|B_n^T(\varphi_n(\theta_n^t) \odot \text{ad}[\varphi_n](\theta_n^t, c_n))\|_\infty \leq O(n^{o(1)}) \forall t \in [T_n]$. Furthermore, from lemma 3(c), it follows that

$$\begin{aligned} |f'(c_n \varphi_n(\theta_n^{t'}))| &\leq (2L^2/\mu)^{1/2} (f(c_n^T \varphi_n(\theta_n^{t'})) - f^*)^{1/2} \\ &\leq O(1)(1 - \eta_n \Omega(n^{\alpha+\gamma}))^{t'/2}. \end{aligned}$$

Therefore,

$$\begin{aligned} \|\theta_n^t\|_\infty &\leq \eta_n O(n^{o(1)}) \sum_{t'=0}^{t-1} (1 - \eta_n \Omega(n^{\alpha+\gamma}))^{t'/2} \\ &\leq \frac{\eta_n O(n^{o(1)})}{1 - (1 - \eta_n \Omega(n^{\alpha+\gamma}))^{1/2}} \\ &\leq O(n^{o(1)-\alpha-\gamma}). \end{aligned}$$

This completes the proof of the lemma \square .

B. Analysis of random family of physical design problems

Problem 1 (Restated) Given a cost function $f : \mathbb{R} \rightarrow \mathbb{R}$ being an L -Lipschitz smooth and μ -strongly convex function that is bounded from below, define a family of physical design problems $\mathcal{F} := \{(f, c_n, \varphi_n \equiv (A_n := G_n^{-1}, B_n := I_n, b_n)) : n \in \mathbb{N}\}$ where

- $\|b_n\|_\infty, \|b_n\|_1, \|c_n\|_\infty, \|c_n\|_1 = \Theta(1)$ and,
- $G_n \in \mathbb{R}^{n \times n}$ is a matrix where each entry is independently drawn from the standard normal distribution.

Proposition 1 (Restated) For the problem of state size n , $(f, c_n, \varphi_n \equiv (A_n, B_n, b_n))$, picked from the family of problems defined in problem 1, it is true that

- With probability $1 - 2 \exp(-\varepsilon^2/2)$, $\|A_n^{-1}\|_2 \leq 2\sqrt{n} + \varepsilon$,
- $\mathbb{E}(\|B_n^T(A_n^{-1}b_n \odot A_n^{-T}c_n)\|_2^2) \geq \Omega(n)$,
- $\mathbb{E}(|c_n^T A_n^{-1}b_n|^2) \leq O(1)$,

and consequently, this family of problems on an average satisfies the $1/2, 1/2$ -asymptotic convergence conditions.

Proof:

- This is a standard result in random matrix theory, for e.g. see Ref. [38].
- It follows from straightforward computation that

$$\begin{aligned} & \mathbb{E}(\|B_n^T(A_n^{-1}b_n \odot A_n^{-T}c_n)\|_2^2) \\ &= \mathbb{E}(\|G_n b_n \odot G_n^T c_n\|_2^2) \\ &= n\|b_n\|^2\|c_n\|^2 + \|b_n \odot c_n\|^2. \end{aligned}$$

Noting that $\|b_n\|^2 \geq \|b_n\|_\infty^2 \geq \Omega(1)$ and $\|c_n\|^2 \geq \|c_n\|_\infty^2 \geq \Omega(1)$, the lemma statement follows.

- I note that,

$$\begin{aligned} \mathbb{E}(|c_n^T A_n^{-1}b_n|^2) &= \mathbb{E}(c_n^T G_n b_n b_n^T G_n^T c_n) \\ &= \|c_n\|^2 \|b_n\|^2. \end{aligned}$$

Since $\|b_n\|^2 \leq \|b_n\|_1^2 \leq O(1)$ and $\|c_n\|^2 \leq \|c_n\|_1^2 \leq O(1)$, the lemma statement follows. \square

Problem 2 (Restated) Given a cost function $f : \mathbb{R} \rightarrow \mathbb{R}$ being an L -Lipschitz smooth and μ -strongly convex function that is bounded from below and $\gamma \geq 1/2$, define a family of physical design problems $\{(f, c_n, \varphi_n \equiv (A_n := R_n \text{diag}(s_n) Q_n^T, B_n := I_n, b_n)) : n \in \mathbb{N}\}$ where

- $\|b_n\|_\infty, \|b_n\|_1, \|c_n\|_\infty, \|c_n\|_1 = \Theta(1)$,
- $s_n \in (0, \infty)^n$ with $\|1/s_n\|_\infty \leq O(n^\gamma)$ and $\|1/s_n\| = \Theta(n)$,
- $R_n, Q_n \in \mathbb{R}^{n \times n}$ are drawn uniformly at random from the Haar measure over orthogonal matrices.

Lemma 8 Let $M = R \text{diag}(v) Q^T \in \mathbb{R}^{d \times d}$, where R, Q are orthogonal matrices drawn independently from the Haar random measure over the set of $d \times d$ orthogonal matrices, and $b, c \in \mathbb{R}^d$, then

$$\mathbb{E}((c^T M b)^2) = \frac{1}{d^2} \|v\|^2 \|b\|^2 \|c\|^2.$$

Proof: Provided in appendix A.

Lemma 9 Let $M = R \text{diag}(v) Q^T \in \mathbb{R}^{d \times d}$, where R, Q are orthogonal matrices drawn independently from the Haar random measure over the set of $d \times d$ orthogonal matrices, and $b, c \in \mathbb{R}^d$, then

$$\mathbb{E}(\|M b \odot M^T c\|^2) = \|b\|^2 \|c\|^2 \|v\|^4 \Omega\left(\frac{1}{d^3}\right).$$

Proof: Provided in appendix A.

Proposition 2 (Restated) For the problem of state size n , $(f, c_n, \varphi_n \equiv (A_n, B_n, b_n))$, picked from the family of problems defined in problem 2,

- $\|A_n^{-1}\|_2 \leq O(n^\gamma)$,
- $\mathbb{E}(\|B_n^T(A_n^{-1}b_n \odot A_n^{-T}c_n)\|_2^2) \geq \Omega(n)$,
- $\mathbb{E}(|c_n^T A_n^{-1}b_n|^2) \leq O(1)$,

and consequently, this family of problems on an average satisfies the $1 - \gamma, \gamma$ -asymptotic convergence conditions.

Proof:

- This follows straightforwardly by noting that $\|R_n\| = \|Q_n\| = 1 \forall n \in \mathbb{N}$ and $\|\text{diag}(s_n)^{-1}\| = \|1/s_n\|_\infty \leq O(n^\gamma)$.
- Using lemma 9, it immediately follows that

$$\begin{aligned} & \mathbb{E}(\|B_n^T(A_n^{-1}b_n \odot A_n^{-T}c_n)\|_2^2) \\ & \geq \|b_n\|^2 \|c_n\|^2 \|1/s_n\|^4 \Omega(n^{-3}). \end{aligned}$$

Noting that $\|b_n\|^2 \|c_n\|^2 \geq \|b_n\|_\infty^2 \|c_n\|_\infty^2 \geq \Omega(1)$ and $\|1/s_n\|^4 \geq \Omega(n^4)$, the lemma statement follows.

- Using lemma 8, it follows that

$$\mathbb{E}((c_n^T A_n^{-1}b_n)^2) = \frac{1}{n^2} \|c_n\|^2 \|b_n\|^2 \|1/s_n\|^2.$$

Since $\|c_n\|^2 \|b_n\|^2 \leq \|c_n\|_1^2 \|b_n\|_1^2 \leq O(1)$, and $\|1/s_n\|^2 \leq O(n^2)$, the lemma statement follows. \square

C. Provably non-resonant problems

Lemma 10 Let $\{\varphi_n \equiv (A_n, B_n, b_n) : n \in \mathbb{N}\}$ be a sequence of physical systems which satisfies $\|b_n\|_1 \leq O(1)$, $\|B_n\|_\infty \leq O(1)$, $\|A_n^{-1}\|_{\max} \leq O(n^{o(1)})$ and $\|A^{-1}\|_2 \leq O(n^\gamma)$ for some

$\gamma < 1/2$, then \forall sequences $\{\theta_n \in \text{dom}(\varphi_n) : n \in \mathbb{N}\}, \{c_n \in \mathbb{R}^n : n \in \mathbb{N}\}$ such that $\|\theta_n\|_\infty \leq O(n^{o(1)-1})$ and $\|c_n\|_1 \leq O(1)$, $\|\varphi_n(\theta_n)\|_\infty \leq O(n^{o(1)})$ and $\|\text{ad}[\varphi_n](\theta_n, c_n)\|_\infty \leq O(n^{o(1)})$.

Proof: It follows straightforwardly that for $n \in \mathbb{N}$,

$$\varphi_n(\theta_n) = A_n^{-1}b_n - A_n^{-1}\text{diag}(B_n\theta_n)\varphi_n(\theta_n).$$

Consequently,

$$\begin{aligned} \|\varphi_n(\theta_n)\|_\infty &\leq \|A_n^{-1}b_n\|_\infty + \|A_n^{-1}\text{diag}(B_n\theta_n)\varphi_n(\theta_n)\|_\infty \\ &\leq \|A_n^{-1}\|_{\max}\|b_n\|_1 + \sqrt{n}\|A_n^{-1}\|_2\|B_n\theta_n\|_\infty\|\varphi_n(\theta_n)\|_\infty. \end{aligned}$$

By assumption, $\|A_n^{-1}\|_2 \leq O(n^\gamma)$ for $\gamma < 1/2$, and $\|B_n\theta_n\|_\infty \leq \|B_n\|_\infty\|\theta_n\|_\infty \leq O(n^{o(1)-1})$. Therefore,

$$(1 - O(n^{o(1)+\gamma-1/2}))\|\varphi_n(\theta_n)\|_\infty \leq O(n^{o(1)}),$$

from which it follows that $\|\varphi_n(\theta_n)\| \leq O(n^{o(1)})$. A similar analysis yields $\|\text{ad}[\varphi_n](\theta_n, c_n)\|_\infty \leq O(n^{o(1)})$. \square

Theorem 2 (Restated) *Let $\mathcal{F} := \{(f, c_n, \varphi_n \equiv (A_n, B_n, b_n)) : n \in \mathbb{N}\}$ be a family of physical design problems that satisfies the α, γ -asymptotic convergence conditions (definition 5) with $\alpha > 1/2, \gamma < 3\alpha$, and also satisfies $\|A_n^{-1}\|_{\max} \leq O(n^{o(1)})$, then gradient descent when applied on (f, c_n, φ_n) , if it does not fail, produces a design θ_n^* such that $f(c_n^T \varphi_n(\theta_n^*)) - f^* \leq \varepsilon$ in $T = \Theta(n^{1-2(\alpha-\gamma)} \log(\varepsilon^{-1}))$ steps.*

Proof: This theorem can be proved by repeating the analysis in the proof of theorem 1, and noting from lemma 10 that the assumption that the gradient descent trajectory $\{\theta_n^1, \theta_n^2 \dots\}$ obtained for the problem (f, c_n, φ_n) satisfies $\|\theta_n^k\|_\infty \leq O(n^{\alpha+\gamma})$ implies that the trajectory is also non-resonant. \square

V. CONCLUSION AND OPEN PROBLEMS

In conclusion, this work provides rigorous evidence for local optimization algorithms being efficient at solving physical

design problems. I show that, under some assumptions on the physics of the system, non-resonant physical design problems are efficiently solvable by gradient descent. Furthermore, I also outline random ensembles of physical design problems which are, on an average, efficiently solvable by local optimization algorithms.

This work, while being a first step towards theoretically understanding the complexity of typical physical design problems, leaves several questions open. One question is to better characterize when a physical design problem is resonant — in this analysis, I need to assume that gradient descent avoids resonant designs in order to show that it globally solves the design problem. While theorem 2 makes some progress in this direction, it would be interesting to more carefully analyze the gradient descent trajectory and obtain a set of weaker conditions on the design problems under which gradient descent avoids resonant devices. Another interesting direction would be to study the optimality of gradient descent on specific design problems, or ensembles of design problems, appearing in practical settings using the tools introduced in this paper. Finally, extending the analysis introduced in this paper to other optimization algorithms (such as quasi-Newton methods like BFGS, L-BFGS, or method of moving asymptotes), would also go a long way in making the rigorous results practically relevant.

ACKNOWLEDGMENTS

I thank Shivam Garg, Logan Su, Geunho Ahn and Alex White for useful discussion. I acknowledge support from Max Planck Harvard Research Center for Quantum Optics (MPHQ) postdoctoral fellowship.

-
- [1] S. Molesky, Z. Lin, A. Y. Piggott, W. Jin, J. Vucković, and A. W. Rodriguez, *Nature Photonics* **12**, 659 (2018).
 - [2] L. Su, A. Y. Piggott, N. V. Saprà, J. Petykiewicz, and J. Vucković, *Acs Photonics* **5**, 301 (2018).
 - [3] A. Y. Piggott, E. Y. Ma, L. Su, G. H. Ahn, N. V. Saprà, D. J. Vercruyse, A. M. Netherton, A. S. Khope, J. E. Bowers, and J. Vucković, arXiv preprint arXiv:1911.03535 (2019).
 - [4] L. Su, R. Trivedi, N. V. Saprà, A. Y. Piggott, D. Vercruyse, and J. Vucković, *Optics express* **26**, 4023 (2018).
 - [5] A. Y. Piggott, J. Petykiewicz, L. Su, and J. Vucković, *Scientific reports* **7**, 1 (2017).
 - [6] N. V. Saprà, D. Vercruyse, L. Su, K. Y. Yang, J. Skarda, A. Y. Piggott, and J. Vucković, *IEEE Journal of Selected Topics in Quantum Electronics* **25**, 1 (2019).
 - [7] J. Roslund and H. Rabitz, *Physical Review A* **79**, 053417 (2009).
 - [8] D. Lucarelli, *Physical Review A* **97**, 062346 (2018).
 - [9] J. Werschnik and E. Gross, *Journal of Physics B: Atomic, Molecular and Optical Physics* **40**, R175 (2007).
 - [10] J. Li, X. Yang, X. Peng, and C.-P. Sun, *Physical review letters* **118**, 150503 (2017).
 - [11] W. Zhu and H. Rabitz, *The Journal of Chemical Physics* **109**, 385 (1998).
 - [12] G. I. Rozvany, *Structural and multidisciplinary optimization* **37**,

- 217 (2009).
- [13] M. Y. Wang, X. Wang, and D. Guo, Computer methods in applied mechanics and engineering **192**, 227 (2003).
- [14] B. Hassani and E. Hinton, *Homogenization and structural topology optimization: theory, practice and software* (Springer Science & Business Media, 2012).
- [15] M. B. Giles and N. A. Pierce, Flow, turbulence and combustion **65**, 393 (2000).
- [16] M. H. Bakr and N. K. Nikolova, IEEE Transactions on Microwave Theory and Techniques **52**, 554 (2004).
- [17] I.-h. Park, I.-G. Kwak, H.-B. Lee, S.-Y. Hahn, and K.-S. Lee, IEEE transactions on magnetics **32**, 1242 (1996).
- [18] J. Dong, K. K. Choi, and N. H. Kim, J. Mech. Des. **126**, 527 (2004).
- [19] E. A. Soliman, M. H. Bakr, and N. K. Nikolova, IEEE transactions on microwave theory and techniques **52**, 589 (2004).
- [20] S. Ruder, arXiv preprint arXiv:1609.04747 (2016).
- [21] P. E. Gill and W. Murray, IMA Journal of Applied Mathematics **9**, 91 (1972).
- [22] O. D. Miller, A. G. Polimeridis, M. H. Reid, C. W. Hsu, B. G. DeLacy, J. D. Joannopoulos, M. Soljačić, and S. G. Johnson, Optics express **24**, 3329 (2016).
- [23] H. Shim, L. Fan, S. G. Johnson, and O. D. Miller, Physical Review X **9**, 011043 (2019).
- [24] P. S. Venkataram, S. Molesky, W. Jin, and A. W. Rodriguez, Physical Review Letters **124**, 013904 (2020).
- [25] S. Molesky, P. S. Venkataram, W. Jin, and A. W. Rodriguez, Physical Review B **101**, 035408 (2020).
- [26] S. Molesky, W. Jin, P. S. Venkataram, and A. W. Rodriguez, Physical Review Letters **123**, 257401 (2019).
- [27] G. Angeris, J. Vuckovic, and S. P. Boyd, ACS Photonics **6**, 1232 (2019).
- [28] Z. Kuang, L. Zhang, and O. D. Miller, Optica **7**, 1746 (2020).
- [29] Z. Kuang and O. D. Miller, Physical Review Letters **125**, 263607 (2020).
- [30] H. Zhang, Z. Kuang, S. Puri, and O. D. Miller, Physical Review Letters **127**, 110506 (2021).
- [31] L. Zhang and O. D. Miller, ACS Photonics **7**, 3116 (2020).
- [32] C. Kern, O. D. Miller, and G. W. Milton, Physical Review Applied **14**, 054068 (2020).
- [33] Z. Allen-Zhu, Y. Li, and Z. Song, in *International Conference on Machine Learning* (PMLR, 2019) pp. 242–252.
- [34] Z. Allen-Zhu, Y. Li, and Y. Liang, Advances in neural information processing systems (2019).
- [35] A. Jacot, F. Gabriel, and C. Hongler, in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (2021) pp. 6–6.
- [36] M. Hardt, T. Ma, and B. Recht, Journal of Machine Learning Research **19**, 1 (2018).
- [37] E. Kogan and M. Kaveh, Physical Review B **52**, R3813 (1995).
- [38] K. R. Davidson and S. J. Szarek, Handbook of the geometry of Banach spaces **1**, 131 (2001).
- [39] B. Collins and P. Śniady, Communications in Mathematical Physics **264**, 773 (2006).
- [40] B. Collins and S. Matsumoto, Journal of Mathematical Physics **50**, 113516 (2009).

Appendix A: Proof of lemmas

This proof requires a basic integration formula with respect to the Haar measure over the orthogonal matrices. Given that $O \in \mathbb{R}^{n \times n}$ is an orthogonal matrix drawn randomly from the Haar measure, this integration formula allows us to express $E(O_{i_1, j_1} O_{i_2, j_2} \dots O_{i_k, j_k})$ in terms of the orthogonal Weingarten function [39, 40]. This function is difficult to evaluate in closed form for a general k , but for my purposes $k = 2$ will suffice. For completeness, I provide this integration formula, and specialize it to $k = 2$ and then provide a proof of proposition 2(b).

Definition 7 (Pairing) For $n \in \mathbb{N}$, denoted by \mathcal{P}_{2n} , is an unordered tuple of n unordered tuples with two elements, $((i_1, j_1), (i_2, j_2) \dots (i_n, j_n))$ where all $i_1, i_2 \dots i_n, j_1, j_2 \dots j_n$ are distinct and $\in [2n]$. The set of all pairings over $[2n]$ will be denoted by \mathcal{P}_{2n} .

Remarks:

- It is important to emphasize the simply reordering the tuples in a pairing, or the the two elements inside the tuple, does not generate a different pairing. For instance, the pairing $((1, 2), (3, 4), (5, 6))$ over $[6]$ is the same as the pairing $((3, 4), (2, 1), (6, 5))$.
- As an explicit and important example for the following calculations, the set \mathcal{P}_4 has three distinct elements, $((1, 2), (3, 4)), ((1, 3), (2, 4)), ((1, 4), (2, 3))$.

Definition 8 (Pairing delta function $\Delta_{i_1, i_2 \dots i_{2n}}^p$) Given a pairing $p = ((k_1, r_1), (k_2, r_2) \dots (k_n, r_n)) \in \mathcal{P}_{2n}$, and indices $i_1, i_2 \dots i_{2n}$ from some set \mathcal{I} , then

$$\Delta_{i_1, i_2 \dots i_{2n}}^p = \prod_{s=1}^n \delta_{i_{k_s}, i_{r_s}},$$

where $\delta_{i,j} = 1$ if $i = j$ and 0 if $i \neq j$ is the Kronecker delta function.

Definition 9 (Loop function $\text{loop}(p_1, p_2)$) Let $p_1, p_2 \in \mathcal{P}_{2k}$ be two pairings. Construct a graph with vertices $\{1, 2, 3 \dots 2k\}$ with edges on all the pairings in p_1 or p_2 , then $\text{loop}(p_1, p_2)$ is the number of connected components in the graph.

Definition 10 (Orthogonal weingarten function $\text{Wg}_k^{O(d)}(p_1, p_2)$) Given $\mathcal{P}_{2k} = \{p_1, p_2, \dots\}$, let $G_k^d \in \mathbb{R}^{|\mathcal{P}_{2k}| \times |\mathcal{P}_{2k}|}$ be a matrix with elements $(G_k^d)_{i,j} = d^{\text{loop}(p_i, p_j)}$, then the $\text{Wg}_k^{O(d)}(p_i, p_j) = [(G_k^d)^{-1}]_{i,j}$.

Lemma 11 For $d > 1$ and for $p_1, p_2 \in \mathcal{P}_4$,

$$\text{Wg}_2^{O(d)}(p_1, p_2) = \begin{cases} \frac{1}{d^2-1} & \text{if } p_1 = p_2, \\ -\frac{1}{d(d^2-1)} & \text{if } p_1 \neq p_2. \end{cases}$$

Proof: Explicitly, $\mathcal{P}_4 = \{p_1 := ((1, 2), (3, 4)), p_2 := ((1, 3), (2, 4)), p_3 := ((1, 4), (2, 3))\}$. I can now calculate the loop function, and it follows that $\text{loop}(p_i, p_j) = 2$ if $i = j$ else 1 and thus

$$G_2^d = \begin{bmatrix} d^2 & d & d \\ d & d^2 & d \\ d & d & d^2 \end{bmatrix} \implies (G_2^d)^{-1} = \frac{1}{d^2-1} \begin{bmatrix} 1 & -1/d & -1/d \\ -1/d & 1 & -1/d \\ -1/d & -1/d & 1 \end{bmatrix}.$$

Identifying the weingarten function with these matrix elements, the lemma statement follows \square .

Lemma 12 (Integration w.r.t. Haar measure over $O(d)$ from Ref. [40]) Let R be a matrix drawn uniformly at random from the Haar measure over the set of $d \times d$ orthogonal matrices, then for all $k \in \mathbb{N}$,

$$\mathbb{E}[R_{i_1, j_1} R_{i_2, j_2}, R_{i_3, j_3} \dots R_{i_{2k}, j_{2k}}] = \sum_{p_1, p_2 \in \mathcal{P}_{2k}} \text{Wg}_k^{O(d)}(p_1, p_2) \Delta_{i_1, i_2, \dots, i_{2k}}^{p_1} \Delta_{j_1, j_2, \dots, j_{2k}}^{p_2}.$$

Lemma 8 (Restated) Let $M = R \text{diag}(v) Q^T \in \mathbb{R}^{d \times d}$, where R, Q are orthogonal matrices drawn independently from the Haar random measure over the set of $d \times d$ orthogonal matrices, and $b, c \in \mathbb{R}^d$, then

$$\mathbb{E}((c^T M b)^2) = \frac{1}{d^2} \|v\|^2 \|b\|^2 \|c\|^2.$$

Proof: Explicitly writing out the expectation value,

$$\mathbb{E}((c^T M b)^2) = \mathbb{E}\left(\sum_{i,j,k \in [d]^2} \prod_{l=1}^2 v_{k_l} c_{i_l} b_{j_l} R_{i_l, k_l} Q_{j_l, k_2}\right) = \sum_{i,j,k \in [d]^2} \left(\prod_{l=1}^2 v_{k_l} c_{i_l} b_{j_l}\right) \mathbb{E}(R_{i_1, k_1} R_{i_2, k_2}) \mathbb{E}(Q_{j_1, k_1} Q_{j_2, k_2}).$$

From lemma 12, it follows that

$$\mathbb{E}(R_{i_1, k_1} R_{i_2, k_2}) = \frac{1}{d} \delta_{i_1, i_2} \delta_{k_1, k_2}, \text{ and } \mathbb{E}(Q_{j_1, k_1} Q_{j_2, k_2}) = \frac{1}{d} \delta_{j_1, j_2} \delta_{k_1, k_2},$$

and consequently,

$$\mathbb{E}((c^T M b)^2) = \frac{1}{d^2} \|v\|^2 \|b\|^2 \|c\|^2.$$

and thus the lemma statement follows \square .

Lemma 9 (Restated) Let $M = R \text{diag}(v) Q^T \in \mathbb{R}^{d \times d}$, where R, Q are orthogonal matrices drawn independently from the Haar random measure over the set of $d \times d$ orthogonal matrices, and $b, c \in \mathbb{R}^d$, then

$$\mathbb{E}(\|M b \odot M^T c\|^2) = \|b\|^2 \|c\|^2 \|v\|^4 \Omega(d^{-3}).$$

Proof: It is easily follows that

$$\mathbb{E}(\|M b \odot M^T c\|^2) = \sum_{i \in [n]} \sum_{j, k, l, m \in [n]^2} v_{j_1} v_{j_2} v_{k_1} v_{k_2} \mathbb{E}(R_{i, j_1} R_{i, j_2} R_{l_1, k_1} R_{l_2, k_2}) \mathbb{E}(Q_{m_1, j_1} Q_{m_2, j_2} Q_{i, k_1} Q_{i, k_2}) b_{m_1} b_{m_2} c_{l_1} c_{l_2}.$$

From lemma 12, it follows that

$$\begin{aligned} \mathbb{E}(R_{i,j_1} R_{i,j_2} R_{l_1,k_1} R_{l_2,k_2}) &= \sum_{p_1, p_2 \in \mathcal{P}_4} \mathbf{Wg}_4^{O(d)}(p_1, p_2) \Delta_{i,i,l_1,l_2}^{p_1} \Delta_{j_1,j_2,k_1,k_2}^{p_2}, \\ \mathbb{E}(Q_{m_1,j_1} Q_{m_2,j_2} Q_{i,k_1} Q_{i,k_2}) &= \sum_{p'_1, p'_2 \in \mathcal{P}_4} \mathbf{Wg}_4^{O(d)}(p'_1, p'_2) \Delta_{m_1,m_2,i,i}^{p'_1} \Delta_{k_1,k_2,j_1,j_2}^{p'_2}. \end{aligned}$$

I first evaluate the summation

$$f(p_1, p'_1, v) := \sum_{j,k \in [d]^2} \sum_{p_2, p'_2 \in \mathcal{P}_4} v_{j_1} v_{j_2} v_{k_1} v_{k_2} \mathbf{Wg}_4^{O(d)}(p_1, p_2) \mathbf{Wg}_4^{O(d)}(p'_1, p'_2) \Delta_{j_1,j_2,k_1,k_2}^{p_2} \Delta_{j_1,j_2,k_1,k_2}^{p'_2}.$$

I note from definition 8

$$\sum_{j,k \in [d]^2} v_{j_1} v_{j_2} v_{k_1} v_{k_2} \Delta_{j_1,j_2,k_1,k_2}^{p_2} \Delta_{k_1,k_2,j_1,j_2}^{p'_2} = \begin{cases} \|v\|_4^4 & \text{if } p_2 \neq p'_2, \\ \|v\|_4^4 & \text{if } p_2 = p'_2. \end{cases}$$

Consequently, it follows that

$$f(p_1, p'_1, v) = (\|v\|^4 - \|v\|_4^4) \sum_{p \in \mathcal{P}_4} \mathbf{Wg}_4^{O(d)}(p_1, p) \mathbf{Wg}_4^{O(d)}(p'_1, p) + \|v\|_4^4 \left(\sum_{p \in \mathcal{P}_4} \mathbf{Wg}_4^{O(d)}(p_1, p) \right) \left(\sum_{p \in \mathcal{P}_4} \mathbf{Wg}_4^{O(d)}(p'_1, p) \right).$$

Using the explicit formula for the Weingarten functions from lemma 11, it follows that

$$\chi(p_1, p'_1, v) = \begin{cases} \chi_{\text{eq}}(v) & \text{if } p_1 = p'_1, \\ \chi_{\text{ueq}}(v) & \text{if } p_1 \neq p'_1. \end{cases}$$

where

$$\begin{aligned} \chi_{\text{eq}}(v) &= \frac{d^2 + 2}{d^2(d^2 - 1)^2} \|v\|^4 - \frac{4d - 2}{d^2(d^2 - 1)^2} \|v\|_4^4, \\ \chi_{\text{ueq}}(v) &= -\frac{2d - 1}{d^2(d^2 - 1)^2} \|v\|^4 + \frac{d^2 - 2d + 3}{d^2(d^2 - 1)^2} \|v\|_4^4 \end{aligned}$$

I note that

$$\begin{aligned} \mathbb{E}(\|Mb \odot M^T c\|^2) &= \sum_{p_1, p'_1 \in \mathcal{P}_4} \sum_{\substack{i \in [n], \\ l, m \in [n]^2}} b_{m_1} b_{m_2} c_{l_1} c_{l_2} \Delta_{i,i,l_1,l_2}^{p_1} \Delta_{m_1,m_2,i,i}^{p'_1} \chi(p_1, p'_1, v) \\ &= \chi_{\text{eq}}(v) (d \|b\|^2 \|c\|^2 + 2 \|b \odot c\|^2) + 6 \chi_{\text{ueq}}(v) \|b\|^2 \|c\|^2. \end{aligned}$$

Using the expressions for $\chi_{\text{eq}}(v)$, $\chi_{\text{ueq}}(v)$, we prove the lemma statement. \square .