Padé Partial Dual Method for Finding Bounds in Inverse Design

Justin Cardona
Engineering Physics Department, Polytechnique Montréal

This work presents a method for forming Lagrange Dual for Inverse Design programs that circumvents the need to check for positive definiteness during optimization. By using a Partial Dual approach and leveraging compact constraints, this problem is replaced with a root finding task that is typically solved here in three iterations using Padé approximations, regardless of system size. The solution is used to efficiently shift the Lagrangian of the system to positive definiteness.

I. INTRODUCTION

Inverse design is the idea of automating the engineering process to algorithmically find designs that outperform human intuition. There have been several examples of this approach surpassing the state of the art for many important applications such as radiative Purcell enhancement [6], metamaterial photonics [11], solar energy [3], and non-linear switching [2] for example. This is typically done by phrasing the creation of an optical device as a mathematical optimization problem and finding its solution. The issue is that these problems are non-convex in general so little inverse design literature can make claims about the global optimality of the designs that are produced.

In addition to this uncertainty, current designs are restricted to very few degrees of freedom. These are highly specialized problems that have high degrees of symmetry or are not even three dimensional, typically having low resolution. There have been some 3D results, but only for sub-wavelength problems (see [14] for example). This is due to the fact that many of the numerical tasks to be performed scale poorly with the degrees of freedom.

A. Electromagnetics Notation

In order to discuss the goal of this thesis more fully, a framework to describe electromagnetics will first be presented. Consider the set of square integrable complex fields in 3D real space $L^2(\mathbb{R}^3, \mathbb{C}^6)$ with the inner product:

$$\langle F|G\rangle = \int_{\mathbb{R}^3} d^3x F(x)^* \cdot G(x) \qquad \forall \left|F\right\rangle, \left|G\right\rangle \in L^2(\mathbb{R}^3, \mathbb{C}^6) \tag{1}$$

Using this, one may write Maxwell's equations as:

$$\begin{aligned} \mu_0 \partial_t(\left|H\right\rangle + \left|m\right\rangle) + \boldsymbol{\nabla} \times \left|E\right\rangle = 0 \\ \epsilon \partial_t \left|E\right\rangle - \boldsymbol{\nabla} \times \left|H\right\rangle = -\left|J\right\rangle. \end{aligned} \tag{2}$$

Fourier transforming in time, they can be written in matrix form as:

$$\begin{pmatrix} \left|J\right\rangle \\ -\left|m\right\rangle \end{pmatrix} = \begin{pmatrix} -i\epsilon\omega & \boldsymbol{\nabla}\times \\ \frac{i}{\mu_0\omega}\boldsymbol{\nabla}\times & 1 \end{pmatrix} \begin{pmatrix} \left|E\right\rangle \\ \left|H\right\rangle \end{pmatrix}. \tag{3}$$

Using magnetic currents instead allows for a cleaner representation and a hermitian matrix. Letting $\left|M\right\rangle = \mu_0 \partial_t \left|m\right\rangle$, $Z = \sqrt{\frac{\mu_0}{\epsilon}}$, and $k = \sqrt{\mu_0 \epsilon}$:

$$\frac{i}{k} \begin{pmatrix} \left|J\right\rangle \\ -\left|M\right\rangle \end{pmatrix} = -\begin{pmatrix} Z^{-1} & -\frac{i}{k}\boldsymbol{\nabla}\times \\ \frac{i}{k}\boldsymbol{\nabla}\times & Z \end{pmatrix} \begin{pmatrix} \left|E\right\rangle \\ \left|H\right\rangle \end{pmatrix}. \tag{4}$$

Concatenating the currents and field together, the above expression may be written as $\frac{i}{k}\left|p\right\rangle = M\left|f\right\rangle$. Additionally, M_0, in place of M, will refer to the Maxwell operator of *free space.* The permeability and permittivity response function will be denoted $X \in \mathrm{GL}\big(L^2(\mathbb{R}^3, \mathbb{C}^6)\big)$, breaking this up into the electric and magnetic components, this relates currents with fields in the typical sense, $\frac{i}{k}\left|p^s\right\rangle = -X\left|f\right\rangle$, or

$$\frac{i}{k} \begin{pmatrix} \left|J^s\right\rangle \\ -\left|M\right\rangle \end{pmatrix} = -\begin{pmatrix} z^{-1}X_{JE} & X_{JH} \\ z^{-1}X_{ME} & X_{MH} \end{pmatrix} \begin{pmatrix} \left|E\right\rangle \\ \left|H\right\rangle \end{pmatrix}. \tag{5}$$

Here, $z = \sqrt{\mu_0/\epsilon_0}$ while i and s superscripts denote incident and scattered fields so that any total field may be written $g = g^i + g^s$. This allows Maxwell's equations in a medium to be written as

$$-\frac{i}{k}\left|p^i\right\rangle = (M_0 - X)\left|f\right\rangle. \tag{6}$$

Instead opting to use the Green's function operator (for free space) G_0, Maxwell's equations are:

$$-\frac{i}{k}\big(X^{-1} - G_0\big)\left|p^s\right\rangle = \left|f^i\right\rangle. \tag{7}$$

Denoting the source field as $\left|S\right\rangle = ik\left|f^i\right\rangle$ and the transmitted currents $\left|T\right\rangle = \left|p^s\right\rangle$, Maxwell's equations imply:

$$\langle S|T\rangle = \langle T|U|T\rangle. \tag{8}$$

with $U = X^{-\dagger} - G_0^\dagger$. The idea behind using currents rather than fields is that it, combined with an integral formalism presents better numerical convergence [4].

B. Inverse Design Context

The goal of inverse design is to automatedly create optical devices. This is done by phrasing the design process as a mathematical optimization problem with an objective function that captures the goal of the project (to be maximized). The design variable is the optical device itself. In other words, one finds the best way of arranging material in physical space in order to achieve the best possible outcome. Mathematically this is the program

$$\underset{X \in \mathrm{GL}(L^2(\mathbb{R}^3,\mathbb{C}^6))}{\text{maximize}} \; f(X) \tag{9}$$

where $f : \mathrm{GL}\big(L^2(\mathbb{R}^3, \mathbb{C}^6)\big) \to \mathbb{R}$. Many important problems in photonics can be represented by net-power-transfer objectives [5]. These take on the following form

$$\underset{X \in \mathrm{GL}(L^2(\mathbb{R}^3,\mathbb{C}^6))}{\text{maximize}} \; \mathrm{Im} \langle S|T\rangle - \langle T|Q|T\rangle \tag{10}$$

where Q is introduced for a general quadratic form. For example, $Q = \mathrm{ASym}G_0$ for absorption/material loss problems, $\mathrm{ASym}X^{-\dagger}$ for scattering/radiation, or 0 for extracted power[5]. In these problems, one is given a source field $|S\rangle$ and an X is found such that the transmitted current $|T\rangle$ has the desires properties. It should be noted that in practice materials have highly localized response, so X is nearly diagonal and is approximated as such. This allows for an easy invertible map between X and $|T\rangle$, namely $|T\rangle = X(|S\rangle, G_0 |T\rangle)$. Therefore, given $|S\rangle$ and $|T\rangle$, X can be found via element-wise division. This greatly simplifies the problem because now one can have a fixed X and treat $|T\rangle$ as the optimization variable in a quadratic program. The tradeoff here is that Maxwell's equations must now be enforced explicitly through equation 8. This problem will be addressed along with discretization.

When actually computing a design, one does not solve for the physical field but rather a discretization of it. For a design region $\mathcal{R}$, one may consider a collection of subsets R. The elements of R are convex and span $\mathcal{R}$. These cells usually lie on a cubic lattice, but for generality an abstract R will do fine. Consider the discretization operator $\delta : L^2(\mathbb{R}^3, \mathbb{C}^6) \to \big(R \to \mathbb{C}^6\big)$ such that $(\delta\,|f\rangle)(r) = |f\rangle\,(\bar{r})$. This is well defined because the centroid $\bar{r} \in \mathcal{R} \subset \mathbb{R}^3$. Similarly for operators $\Delta : \mathrm{GL}\big(L^2(\mathbb{R}^3, \mathbb{C}^6)\big) \to \mathrm{GL}\big(R \to \mathbb{C}^6\big)$, the evaluation happens as $(\Delta G)(|f\rangle)(r) = G\,|f\rangle\,(\bar{r})$. Now for a final note on how to enforce Maxwell's equations. The simplest way is to say that $\forall r \in R$ equation 8 holds. However, for computational efficiency, it might be sufficient to consider a partitioning set of connected subsections of R, Ω_R and impose [6]:

$$\forall \omega \in \Omega_R \; \langle S|\mathbb{I}_\omega|T\rangle = \langle T|U\mathbb{I}_\omega|T\rangle \tag{11}$$

With this in mind, note that further discussion will occur in the discretized context. That is to say that there is a change of variables occuring: $|T\rangle \to \delta\,|T\rangle$, $Q \to \Delta Q$, etc. The fully discretized optimization problem may then be written as

$$\begin{aligned} \mathrm{P} = \underset{|T\rangle \in R \to \mathbb{C}^6}{\text{maximize}} \quad & \mathrm{Im} \langle S|T\rangle - \langle T|Q|T\rangle \\ \text{such that } \forall \omega \in \Omega_R \; & \langle S|\mathbb{I}_\omega|T\rangle = \langle T|U\mathbb{I}_\omega|T\rangle \end{aligned} \tag{12}$$

The goal is ultimately to efficiently find globally optimal solutions to the above program when the number of elements in R is very large. In general, it may be difficult to find the global optima as the problem is not convex. However, by using Lagrange duality it is always possible to find a convex optimization problem that yields an upper bound on design performance. Therefore it is very important to be able to solve this problem.

C. Lagrange Duality

In order to have further insight into this problem, an overview of some key ideas of Lagrange duality will be presented. Consider the following program:

$$\begin{aligned} \mathrm{P} = \underset{x \in \mathbb{C}^n}{\text{maximize}} \; & f_0(x) \\ \text{such that } \forall m \in \{1, \dots M\} & f_m \geq 0. \end{aligned} \tag{13}$$

where $\forall m \in \{0, \dots M\}$ $f_m : \mathbb{C}^n \to \mathbb{R}$. This can alternatively be represented as an unconstrained optimization through the use of indicator functions.

$$\mathrm{P} = \underset{x \in \mathbb{C}^n}{\text{maximize}} \; f_0(x) + \sum_{m=1}^{M} \mathbb{I}(f_m(x) \geq 0) \tag{14}$$

where the indicators are defined as:

$$\mathbb{I}(x) = \begin{cases} 0 & \text{if } x \text{ is true} \\ -\infty & \text{if } x \text{ is false} \end{cases} \tag{15}$$

The discontinuous nature of this objective makes it impractical to use gradient methods to find a solution. Therefore, instead of using indicators one can use a linear barrier function with positive coefficients. This will reward a positive value for the constraint function while penalizing negative values.

$$\underset{x \in \mathbb{C}^n}{\text{maximize}} \; f_0(x) + \sum_{m=1}^{M} \lambda_m f_m(x) \tag{16}$$

The tradeoff for this continuous form is that the reward terms now add a non-zero portion to the objective. However, if x is in the feasible set F_P (it satisfies the constraints), then this is still an upper bound.

$\mathcal{L}(x,\lambda) = f_0(x)+\sum_{m=1}^{M} \lambda_m f_m(x)$ is referred to as the Lagrangian, and $\mathcal{D}(\lambda) = \sup_{x\in\mathbb{C}^N} \mathcal{L}(x,\lambda)$ is the dual function. In order to find the best lower bound, one can minimize the dual function. This results in the dual program:

$$\mathrm{D(P)} = \inf_{\lambda\in\mathbb{R}^M_{\geq}} \sup_{x\in\mathbb{C}^N} \mathcal{L}(x,\lambda) \tag{17}$$

Note that this problem is convex in the Lagrange multipliers. Furthermore, note that for any feasible x, $\mathcal{L}$ is an upper bound to f_0. Since the constraints are satisfied, the f_m terms can only increase the value of the function. The dual approach is valuable for this reason: it is a convex problem that bounds the primal one. Therefore standard optimization methods can be used to guarantee a bound on the performance of any design.

II. DUALITY IN NANOPHOTONICS

Returning to equation 12, for the sake of the computational context, instead of working on $R \to \mathbb{C}^6$, since R is finite, the space $\mathbb{C}^N$ will be used with $N = 6|R|$. Similarly $Q, U, \ldots$ can be represented by matrices in $\mathbb{C}^{N\times N}$.

$$\begin{aligned} \mathrm{P} = \underset{|T\rangle\in\mathbb{C}^N}{\text{maximize}} \quad & \mathrm{Im}\,\langle S|T\rangle - \langle T|Q|T\rangle \\ \text{such that } \forall\omega\in\Omega_R \; & \langle S|\mathbb{I}_\omega|T\rangle = \langle T|U\mathbb{I}_\omega|T\rangle \end{aligned} \tag{18}$$

One may note that the Hermitian and anti-Hermitian parts of the equation hold independently and thus can be separated into two constraints. The tendency of this optimization is to increase the magnitude of $|T\rangle$. Therefore the only side of the equalities that ends up being enforced are the "$\geq$" ones as they provide an upper bound to the norm of $|T\rangle$:

$$\begin{aligned} \mathrm{P} = \underset{|T\rangle\in\mathbb{C}^N}{\text{maximize}}\, & \mathrm{Im}\,\langle S|T\rangle - \langle T|Q|T\rangle \\ \text{such that } \forall\omega\in\Omega_R\; & \mathrm{Im}\,\langle S|\mathbb{I}_\omega|T\rangle \geq \langle T|\mathrm{Asym}(U\mathbb{I}_\omega)|T\rangle \\ & \mathrm{Re}\,\langle S|\mathbb{I}_\omega|T\rangle \geq \langle T|\mathrm{Sym}(U\mathbb{I}_\omega)|T\rangle \end{aligned} \tag{19}$$

Now that the problem has been properly stated, it is time to begin finding the dual problem. The goal of this is to obtain a convex optimization problem that can be solved using standard gradient methods in order to give a bound on the primal. The Lagrangian of this program is:

$$\begin{aligned} \mathcal{L} = {} & \mathrm{Im}\,\langle S|T\rangle - \langle T|Q|T\rangle \\ & + \sum_{m=1}^{M} \alpha_m (\mathrm{Re}\,\langle S|\mathbb{I}_\omega|T\rangle - \langle T|\mathrm{Sym}(U\mathbb{I}_\omega)|T\rangle) \\ & + \sum_{m=1}^{M} \beta_m (\mathrm{Im}\,\langle S|\mathbb{I}_\omega|T\rangle - \langle T|\mathrm{ASym}(U\mathbb{I}_\omega)|T\rangle) \end{aligned} \tag{20}$$

but here M$= |\Omega_R|$ and $\mathbb{I}_m$ is the projection onto the m^{th} element of Ω_R.

$$\begin{aligned} \mathcal{L} &= \begin{pmatrix} \langle T| & \langle S| \end{pmatrix} \begin{pmatrix} -L_{TT} & T_{TS} \\ L_{TS}^\dagger & 0 \end{pmatrix} \begin{pmatrix} |T\rangle \\ |S\rangle \end{pmatrix} \\ L_{TT} &= Q + \sum_{m=1}^{M} (\alpha_m \langle T|\mathrm{Sym}(U\mathbb{I}_\omega)|T\rangle \\ &\qquad + \beta_m \langle T|\mathrm{ASym}(U\mathbb{I}_\omega)|T\rangle) \\ L_{TS} &= \frac{1}{2i}\mathbb{I} + \frac{1}{2}\sum_{m=1}^{M} (\alpha_m + i\beta_m)\mathbb{I}_m \end{aligned} \tag{21}$$

In order for the dual to give a meaningful bound on the primal, $L = \begin{pmatrix} -L_{TT} & L_{TS} \\ L_{TS}^\dagger & 0 \end{pmatrix} \succcurlyeq 0$ is needed. This requires that $L_TT \succcurlyeq 0$.

A. Dual Numerical Difficulties

In order to show whether the matrix is positive definite, one of two approaches is typically done. The first is by checking whether the Cholesky decomposition terminates [8, 9]. Given a matrix $A \in \mathbb{C}^{N\times N}$, a Cholesky decomposition is when A can be written as $LL^\dagger$ where L a triangular matrix with real and positive diagonal entries. It is a necessary and sufficient condition [10]. The issue is that the discretization of the design region may contain potentially millions of cells (elements of R), so Cholesky decomposing L_{TT} is completely impractical given its cubic scaling [**?**]. For this reason it is unfeasible, especially in the case of large-scale inverse design.

An alternative approach that is commonly used are randomized trace estimators for log barrier methods [1]. The most performant and generally applicable method is the Hutchinson trace estimator, these work in the following way. Assume it is known that the matrix is initially positive definite (known and by doing a costly method). Each time the Lagrange multipliers are updated, it must be verified whether the matrix is still positive definite. Since the multipliers undergo very small changes, and the eigenvalues have a continuous dependence on them (from the characteristic equation), one might intuitively expect that the eigenvalues of L_{TT} will become vanishing small before they become negative. One can test for this case by evaluating $\log\det L_{TT}$. This is due to the fact that $\det L_{TT} = \prod_{n=1}^{N} \lambda_n$, the product of L_{TT}'s eigenvalues. This means $\log\det L_{TT} = \sum_{n=1}^{N} \log\lambda_n$. Note that when $\lambda_n \to 0$ $\log\lambda_n \to -\infty$ so when $\log\det L_{TT}$ is large and negative, it indicates that L_{TT} is nearly indefinite.

While $\det L_{TT}$ it is very expensive to evaluate, note that

$$\log\det L_{TT} = \sum_{n=1}^{N} \log\lambda_n = \mathrm{Tr}\log L_{TT} \tag{22}$$

so one could estimate the trace instead of the determinant, which is cheaper. First a series expansion is found:

$$\begin{aligned}\log L_{TT} &= \log(I + (L_{TT} - I)) && (23)\\ &= \sum_{k=1}^{\infty} \frac{-1^{k+1}}{k}(L_{TT} - I)^k && (24)\end{aligned}$$

and then the trace is evaluated:

$$\operatorname{Tr}\log L_{TT} = \sum_{k=1}^{\infty} \frac{-1^{k+1}}{k}\operatorname{Tr}(L_{TT} - I)^k \qquad (25)$$

Now, regarding how the trace is efficiently estimated. The Hutchinson trace is defined for $A \in \mathbb{C}^{m\times m}$ as [12]:

$$\operatorname{Tr}_{H^n} A = \frac{1}{n}\sum_{i=1}^{n} \langle z_i|A|z_i\rangle\,, \quad z \sim \{-1, 1\}^m \qquad (26)$$

With probability $1 - \delta$ this has a relative error of ϵ, given that n samples have been drawn:

$$n = \frac{2}{\epsilon^2}\left(2 + \frac{8\sqrt{2}}{3}\right)\log\left(\frac{2}{\delta}\right) \qquad (27)$$

There are a few issues with this approach. Firstly, the log expansion has truncation error that compounds with this. Secondly, the series requires $\|L_{TT} - I\| < 1$ which is not always true. This is usually circumvented by doing $\log A = 2n \log A^{1/2n}$, but square roots are too expensive here. Even assuming complete accuracy in the log, for a 99% chance of a 0.1% error 30,460,939 sample matrix vector products are needed. For a 75% chance of a 1% error 84,746 matrix vector products are needed.

These methods are very expensive a better alternative will be presented. Recall the Lagrangian of the primal problem 19. In order to circumvent having to check whether $L_{TT} \succ 0$, a compact constraint is singled out and imposed explicitly. Namely, this will be the anti-symmetric constraint over the entire domain. This means making the change $R \to R \setminus \{\mathcal{R}\}$ in equation 19 in addition to adding the constraint explicitly. Of course D(P) with $\mathcal{D}$ in this form is still dual to P, but this form allows for another manipulation, finding a *partial* dual $\mathrm{D}^{\partial}(P)$ which has the Lagrangian [6]:

$$\begin{aligned}\mathcal{L}^{\partial} &= \big(\langle T| \ \ \langle S|\big) L \begin{pmatrix}|T\rangle\\ |S\rangle\end{pmatrix} + \gamma(\operatorname{Im}\langle S|T\rangle - \langle T|\mathrm{ASym}U|T\rangle)\\ &= \big(\langle T| \ \ \langle S|\big) L^{\partial} \begin{pmatrix}|T\rangle\\ |S\rangle\end{pmatrix} && (28)\end{aligned}$$

Note that this means $L^{\partial}_{TT} = L_{TT} + \gamma E$, $L^{\partial}_{TS} = L_{TS} + i\gamma\mathbb{I}$, with $E = \mathrm{Asym}U$. The real potential issue here is L_{TT}, but since $E \succ 0$, if γ is large enough L^{∂} is as well. When this is the case, the optimal value of $|T\rangle$ can be expressed in terms of the Lagrange multipliers by finding stationary points:

$$\begin{aligned}\frac{\partial}{\partial\,|T\rangle}\mathcal{L}^{\partial} &= \frac{\partial}{\partial\,|T\rangle}\Big(-\langle T|L^{\partial}_{TT}|T\rangle + \langle T|L^{\partial}_{TS}|S\rangle + \langle S|L^{\partial\dagger}_{TS}|T\rangle\Big)\\ &\implies L^{\partial}_{TT}\,|T\rangle = L^{\partial}_{TS}\,|S\rangle && (29)\end{aligned}$$

The goal therefore, is to find the smallest partial dual multiplier γ_* such that $L_{TT} + \gamma_* E \succcurlyeq 0$ and $C_\gamma = \operatorname{Im}\langle S|T_\gamma\rangle - \langle T_\gamma|E|T_\gamma\rangle \geq 0$. To show that such a solution does exist, consider the derivatives of C_γ:

$$\begin{aligned}\frac{dC_\gamma}{d\gamma} &= 2\left(\frac{1}{2}\langle S| - i\,\langle T_\gamma|\, E\right) L^{\partial-1}_{TT}\left(\frac{1}{2}\,|S\rangle + iE\,|T\rangle\right)\\ \frac{d^2C_\gamma}{d\gamma^2} &= -6\left(\frac{1}{2}\langle S| - i\,\langle T_\gamma|\, E\right) L^{\partial-1}_{TT} E L^{\partial-1}_{TT}\left(\frac{1}{2}\,|S\rangle + iE\,|T\rangle\right)\end{aligned}$$

When $L^{\partial}_{TT} \geq 0$ then both derivatives are positive and:

$$\lim_{\gamma\to\infty} C_\gamma = \frac{1}{4}\langle S|E^{-1}|S\rangle$$

Therefore there does exist γ_* that satisfies the previously stated criteria. To find it such that $L^{\partial}_{TT} \succcurlyeq 0$ one must find the most negative solution to the generalized eigenvalue problem $L_{TT} + \gamma_* E \succcurlyeq 0$. Herein lies the issue, the usual iterative methods for doing this are also computationally expensive. Solving the problem for only the last pole γ_p:

$$\begin{aligned}\lim_{\gamma\to\gamma_p^+} C_\gamma &= -\infty\\ \lim_{\gamma\to\gamma_p^+} \partial_\gamma C_\gamma &= \infty\\ \lim_{\gamma\to\gamma_p^+} \partial^2_\gamma C_\gamma &= -\infty\end{aligned}$$

After this point, however C_γ is increasing, concave, and asymptotes to a constant positive value. Therefore $\gamma_* > \gamma_p$ (intermediate value theorem).

B. Solution with Padé Approximation

Suppose that at the start of the problem the most largest eigenvalue problem solution or last zero crossing is known. This could be either through the previously mentioned expensive methods or otherwise. The workaround uses the fact that in gradient methods, Lagrange multipliers will be perturbed by small amount so that the γ_* should also not change much (C_γ is a continuous function $\forall\gamma > \gamma_p$). The idea is to use the previous γ_* as an initial guess and search close to it. The typical approach to this might be to use a gradient descent such as Newton's method. This is not desirable here since evaluating the derivative can be numerically unstable and requires additional inverse solves. Therefore, a derivative free process is preferred. One might be inclined to use something such as the Secant method or higher order variations. These are all essentially equivalent to fitting the constraint function to a polynomial fit, and then finding the zero of the fit. However, due to the rapid change in behaviour of the function around its zero and the presence of poles it will be better suited to use rational functions instead of polynomials to do the fit[7]. These approximations have an analytic expression for the zero that is quick to evaluate for a small number of sample points. The idea therefore is to evaluate C_γ as few times as possible using these basis functions instead. A Padé approximation is a rational function of the form:

$$r(z) = \frac{n(z)}{d(z)} = \sum_{j=1}^{m} \frac{w_j f_j}{z - z_j} \Bigg/ \sum_{j=1}^{m} \frac{w_j}{z - z_j} \tag{30}$$

The approximant is constructed according to the AAA algorithm. It works as follows: consider a function $f : \mathbb{C} \to \mathbb{C}$, the goal is to find another function $r : \mathbb{C} \to \mathbb{C}$ to approximate it. Given a finite ordered set $Z \subset \mathbb{C}$ and its corresponding f values $F \subset \mathbb{C}$. The AAA algorithm will split Z and F, each into 2 partitions. The first partitions (z^m and f^m) will be used to make the m support point in the Padé series.

The second partitions ($Z^m = Z \backslash z^m$ and $F^m = F \backslash f^m$) will be used as sample points to do a least squares fit the series, thus determining the weights. This is accomplished in the following way:

1. $f^1 = \left\{\text{argmin}_{f \in F}(f - \langle F \rangle)^2\right\}$ and z^1 is the corresponding singleton. This defines F^1 and Z^1.
2. Obtain w^1 by performing a least squares fit to F^1 and Z^1 using f^1 and z^1 as the support point.
3. Given F^m, Z^m, f^m, z^m, and w^m calculate the square residuals for all $z \in Z^m$. Select the $(z, f) \in Z^m \times F^m$ with the least residual and add it to the support points.
4. Using the newly obtained f^{m+1} and z^{m+1}, obtain w^{m+1} by least squares fitting to F^{m+1} and Z^{m+1}.
5. Repeat steps 3 and 4 until the desired error tolerance is reached.

Now for how this is applied to the partial dual:

1. Sample on a uniform distribution centered on an initial guess to form Z and use C_ζ to form F.
2. Form the AAA approximant of C_γ using Z and F and find the zero of the approximant γ_{guess}
3. If $C_{\gamma_{\text{guess}}}$ is not within tolerance, add $(\gamma_{\text{guess}}, C_{\gamma_{\text{guess}}})$ to (Z, F) and return to step 2.

While finding the initial guess might be costly, this must only be done once in the entire inverse design. For subsequent steps during a search in the Lagrange multiplier space, the L matrices are perturbed according to their dependance on λ. Since their dependance is continuous, it is expected that small changes in λ will result in small changes in the last zero crossing.

III. RESULTS

To assess the effectiveness of this method a variety of optimization problems were generated by uniformly randomly choosing media and source fields. The χ of the system was chosen over the support $[-0.5, 0.5] \times [0, i10^{-3}]$ and $|S\rangle$ over the support $[-0.5, 0.5] \times [-0.5i, 0.5i]$. The computational domains consisted 3D cubic lattice discretizations of sizes spanning several orders of magnitude. For each size, 100 such samples were drawn. Each sample experienced 10 perturbations of the lagrange multipliers according to a gaussian random walk ($\mu = 0, \sigma = 10^{-2}$) for a total of 1000 data points per system size.

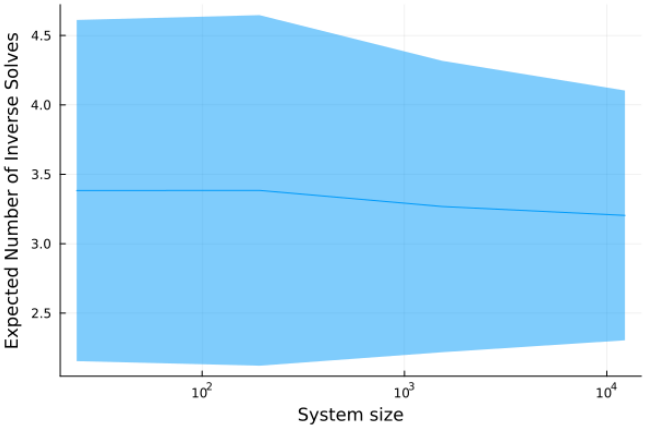

FIG. 1: The number of sample points needed to converge to γ_* to `float32` precision

There is no discernable scaling of required inverse solves with the size of the system (Figure 1) and overall the method requires no more than 3 inverse solve the vast majority of the time (Figure 2).

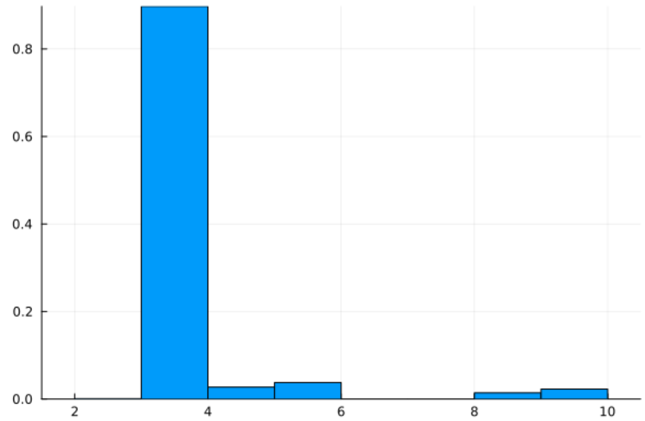

FIG. 2: The distribution of number of sample points needed to converge to γ_* to `float32` precision. This takes on average 3.3 samples with a standard deviation of 1.1.

IV. DISCUSSION

This method's constant scaling in the number of iterations is not the whole story, as the system size increases the computational cost of sampling also increases. The green's operator is stored in a matrix-free fashion so BiCGStab is used to compute solutions [13]. A matrix-vector product here $O(n \log n)$ since GILA and for the largest system size tested here about 20 matrix-vector products are needed for each inverse solve. While a specific comparison might be imprecise as there exist different implementations of Cholesky decomposition, once one goes beyond n of a few hundred it is clear that the Padé based approach is superior.

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] Yuriy Elesin, Boyan Stefanov Lazarov, Jakob Søndergaard Jensen, and Ole Sigmund. Design of robust and efficient photonic switches using topology optimization. *Photonics and nanostructures-Fundamentals and Applications*, 10(1):153–165, 2012.

[3] Vidya Ganapati, Owen D Miller, and Eli Yablonovitch. Light trapping textures designed by electromagnetic optimization for subwavelength thick solar cells. *IEEE Journal of Photovoltaics*, 4(1):175–182, 2013.

[4] Johannes Markkanen, Pasi Yla-Oijala, and Ari Sihvola. Discretization of volume integral equation formulations for extremely anisotropic materials. *IEEE Transactions on Antennas and Propagation*, 60(11):5195–5202, 2012.

[5] Sean Molesky, Pengning Chao, Weiliang Jin, and Alejandro W Rodriguez. Global t operator bounds on electromagnetic scattering: Upper bounds on far-field cross sections. *Physical Review Research*, 2(3):033172, 2020.

[6] Sean Molesky, Pengning Chao, and Alejandro W Rodriguez. Hierarchical mean-field t operator bounds on electromagnetic scattering: Upper bounds on near-field radiative purcell enhancement. *Physical Review Research*, 2(4):043398, 2020.

[7] Yuji Nakatsukasa, Olivier Sète, and Lloyd N Trefethen. The aaa algorithm for rational approximation. *SIAM Journal on Scientific Computing*, 40(3):A1494–A1522, 2018.

[8] Pierre Roux. Formal proofs of rounding error bounds: With application to an automatic positive definiteness check. *Journal of Automated Reasoning*, 57:135–156, 2016.

[9] Siegfried M Rump. Verification of positive definiteness. *BIT Numerical Mathematics*, 46:433–452, 2006.

[10] Robert B Schnabel and Elizabeth Eskow. A new modified cholesky factorization. *SIAM Journal on Scientific and Statistical Computing*, 11(6):1136–1158, 1990.

[11] Bing Shen, Peng Wang, Randy Polson, and Rajesh Menon. Ultra-high-efficiency metamaterial polarizer. *Optica*, 1(5):356–360, 2014.

[12] Maciej Skorski. Modern analysis of hutchinson's trace estimator. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–5. IEEE, 2021.

[13] Henk A Van der Vorst. Bi-cgstab: A fast and smoothly converging variant of bi-cg for the solution of nonsymmetric linear systems. *SIAM Journal on scientific and Statistical Computing*, 13(2):631–644, 1992.

[14] Wenjin Xue, Hanwen Zhang, Abinand Gopal, Vladimir Rokhlin, and Owen D Miller. Fullwave design of cm-scale cylindrical metasurfaces via fast direct solvers. *arXiv preprint arXiv:2308.08569*, 2023.