Chapter 7

From Inverse Problems to Device Design

7.1 3:1 Waveguide Splitter

Encouraged by our analysis in the last section, we now attempt our first simple but non-trivial design problem where we wish to have an incoming waveguide split into 2 branches with a ratio of 3 to 1, and then recombined. We retain the $7a \times 7a$ supercell geometry and construct a mode that complements the chosen hexagonal lattice. The mode is created by first adopting a transverse Gaussian profile along the splitter path, then modulating each segment with an appropriate plane wave. To produce the 3:1 split, we attenuate the upper branch amplitude to 25% of the unbranched intensity, and the lower branch to 75%, as in figure 7.1. We choose the desired frequency ($\omega a/2\pi c$) to be 0.2081, which lies in the middle of the bandgap. We choose the simpler Tikhonov scheme and use the defect-free lattice as a starting point again. Figure 7.2 shows the regularized solution with a residual norm of 3.4491 using a regularization parameter of 32.4. In figure 7.3 we show the L-curve for this problem, where $\lambda_{corner} \approx 1.5$. The much more subtle 'corner' is more typical of a real design problem, and we found (as we mentioned in section 3.3) that the best solution is not near the corner anyway. The solution at $\lambda = 1.5$ is shown in figure 7.4, and again, notice the scale on the colorbar and the substantial amount of noise in the solution.

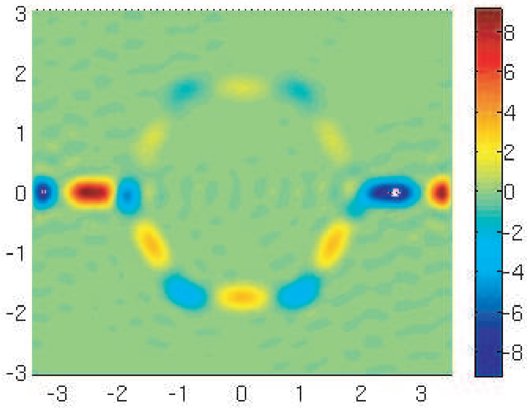

Figure 7.1: Target mode for the 3:1 splitter device.

In general, we try several parameters until a sensible solution is found. From our experience, we often have to regularize beyond the L-curve corner, and ‘interpret’ the result by smoothing out the noisy components. Generally, we look at a particular solution and identify dominant features, and then propagate the smoothed solution through the forward problem again. In this case (ignoring the imaginary parts again), we find the real part of the dielectric suggests a geometry where the upper branch has a hole radius that is half that of the bulk hole radius, and the lower branch has completely filled holes as in figure 7.5. Propagating this geometry through the forward problem yields the 3:1 split mode as shown in figure 7.6. Notice that the actual mode we obtained was not identical to the original target mode, with the obvious difference between the two being that the target mode was more localized than the obtained mode. However, since our design goal was only to construct a 3:1 split waveguide, we can stop here. The discrepancy between the two should not be surprising in light of the magnitude of the residual norm, as we discuss in the following section.

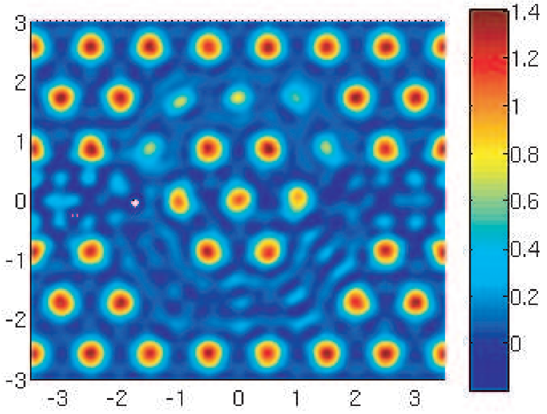

Figure 7.2: Regularized solution of the 3:1 splitter inverse problem using a regularization parameter of 32.4

7.2 Residual Norm

Recall that the expression for our inverse equation comes from the Helmholtz equation:

$$\nabla \times \Big(\eta(\mathbf{r})\nabla \times \mathbf{H_m}(\mathbf{r})\Big) = \frac{\omega^2}{c^2}\mathbf{H_m}(\mathbf{r}) \tag{7.1}$$

$$\therefore \quad \Theta^{(\eta)}h^{(m)} = \frac{\omega^2}{c^2}h^{(m)} \equiv b \tag{7.2}$$

$$A^{h^{(m)}}\eta = b \qquad \text{and,} \tag{7.3}$$

$$A^{h^{(m)}}\eta = \Theta^{(\eta)}h^{(m)} \tag{7.4}$$

$\eta(\mathbf{r})$ and $H_m(\mathbf{r})$ are a special pairing since $\eta(r)$ is the dielectric that supports $H_m(\mathbf{r})$ as an eigenmode. The equality in equation (7.1) only holds if the two are an 'eigenpair'. For some given solution η_{sol} we find in solving the inverse problem for which $A\eta_{sol} \neq b$, it necessarily means that $H_m(\mathbf{r})$ is not an eigenmode of η_{sol}. This finite residual norm is significant for the design problem, because it means we are guaranteed **not** to obtain our desired mode. The obvious question to ask is what exactly are we getting (using a residual norm metric) when we don't find an eigenpair? We can examine the

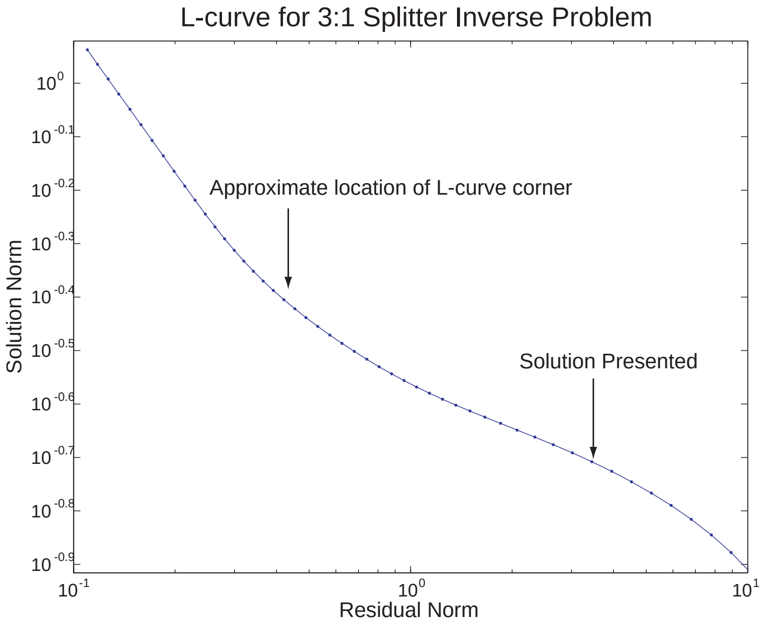

Figure 7.3: L-curve for the 3:1 splitter showing the location and shape of the corner as well as the optimal solution.

left hand side of the equation using η_{sol} to form the Helmholtz operator Θ.

$$A^{h^{(m)}}\eta_{sol} - b \neq 0 \tag{7.5}$$

$$\Theta^{(\eta_{sol})}h^{(m)} - b \neq 0 \tag{7.6}$$

Since the forward problem is well-posed, we can always find the spectrum of eigenmodes for any given dielectric geometry. Assume the following spectral decomposi-

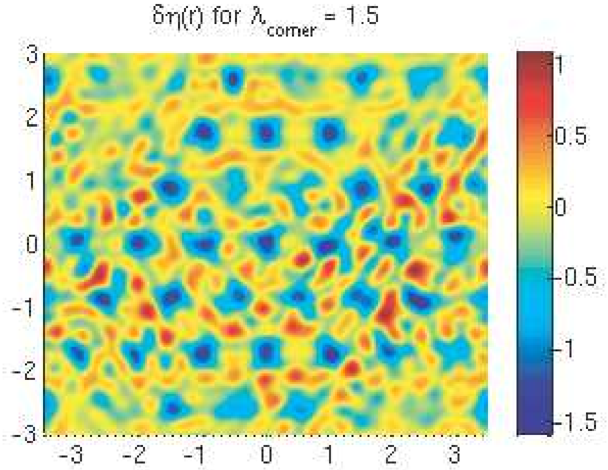

Figure 7.4: 3:1 splitter $\delta\eta(r)$ solution using regularization parameter from the L-curve corner. The information in this 'solution' is practically useless.

tion:

$$
\begin{aligned}
\Theta^{(\eta_{sol})} h_i^{\eta_{sol}} &= \frac{\omega_i^{(\eta)^2}}{c^2} h_i^{\eta_{sol}} \\
\text{Let} \qquad h^{(m)} &\equiv \sum_i \alpha_i h_i^{\eta_{sol}} \\
\text{Then} \qquad \Theta^{(\eta_{sol})} h^{(m)} - b &= \Theta^{(\eta_{sol})} h^{(m)} - \frac{\omega_m^2}{c^2} h^{(m)} \\
&= \Theta^{(\eta_{sol})} \sum_i \alpha_i h_i^{\eta_{sol}} - \frac{\omega_m^2}{c^2} \sum_i \alpha_i h_i^{\eta_{sol}} \\
&= \sum_i \frac{(\omega_i^2 - \omega_m^2)}{c^2} \alpha_i h_i^{\eta_{sol}}
\end{aligned}
\tag{7.7}
$$

Even though minimizing the residual norm over (η_{sol}) may be the 'best' general strategy for solving an ill-conditioned linear system of equations, it becomes clear that it is not the most appropriate strategy for the purpose of the design problem. Using the eigenvalue decomposition shows explicitly that the entire spectrum of eigenmodes of η_{sol} contribute to the residual norm, whereas we only care that our desired mode be close to a single eigenmode of η_{sol}. Furthermore, each eigenmode in the spectrum is component-wise weighted by the ω_i^2 term. Since the frequency of the modes of interest usually lie within the first bandgap (i.e. relatively low frequencies), the residual

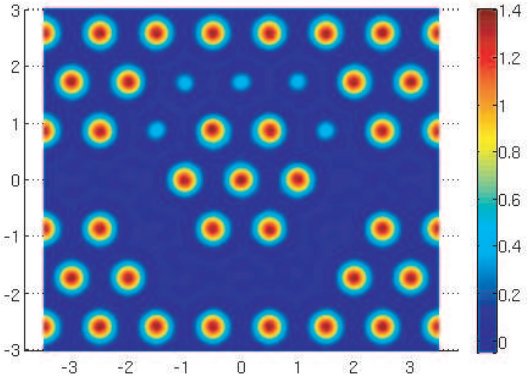

Figure 7.5: Actual $\eta(r)$ used based on the output solution of the inverse problem (see figure 7.2.

norm metric overemphasizes high frequency mode contributions. Therefore, the minimization has a bias towards dielectric geometries whose high frequency eigenmodes are orthogonal to the desired mode at the expense of a stronger overlap of a single eigenmode.

As an illustration, we can repeat our perturbed inverse problem using the noisy h_1 mode as our target mode (only we add an even smaller perturbation ($|n| = 10^{-3} \times |h_m|$) to the mode and renormalize). Rather than looking to solve the equation as we did in section 6.5, we enter our exact solution $\eta_{\mathbf{k}}^{h_1}$ into $|A\eta - b|$ and evaluate the residual norm. Even though it would produce our target mode minus the small bit of noise, the residual is 1.59. Because of the high frequency components, we see that the best solution to the physical problem does not even come close to solving the linear equation.

In situations where the desired mode happens to be an *exact* eigenmode of some geometry, the residual norm metric is fine, because it will simply find the complementary solution (as in our contrived example). The inverse equations find a solution whose spectral decomposition of the target mode is a pure eigenmode, so it can afford to ignore the high frequency mode spectrum. However, this is a rare occurrence, and we will elaborate further in section 7.3. As we deviate further from an exact eigen-

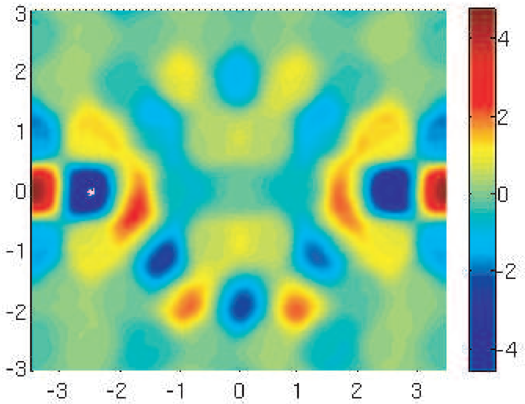

Figure 7.6: Magnetic field distribution of the 3:1 splitter mode of interest supported by the design of Fig. 7.5.

mode, the effect of the errors become more observable, as we saw in section 6.5. In any realistic design problem then, what benefit can we get out of this approach? This is precisely the phenomenon we see with the 3:1 waveguide splitter. We found that the obtained eigenmode was actually quite different from the desired one, and yet it achieved our goal. By examining the residual norm more closely, we now understand that the optimality of our result depends somewhat fortuitously on the target mode's proximity to a feasible eigenmode. The more overlap between the target mode and a feasible eigenmode, the less it needs to compromise with the high frequency modes. However, we also see that despite this non-optimal bias, it still manages to find a 'reasonable' result. Particularly in situations where we truly have little intuition into the problem, this will at least give us a good starting point. Therefore, we cannot claim the resulting mode is **optimal** as we might have hoped, but we find that it is still 'better' than what we may otherwise have.

7.3 Inverse Problem Based Design Flow

We began chapter 6 by motivating an inverse problem based design approach to obtain 'optimal' PBG structures, but so far we have really only outlined a single step: that of solving the inverse problem. In the previous section, we showed how even this one step is often not optimal, and we discovered the importance of a ‘good’ target mode. In this section, we expand on the entire design process, highlighting other difficulties to achieving optimality.

Recall that the idea of optimality is connected to a performance metric that is some function of the desired field mode, and possibly of the dielectric function as well. To find an optimal design implies finding the optimum of the performance function, but there are unfortunately no guarantees that one can actually find the global optimum of such a function. High-dimensional global optimization for arbitrary functions is notoriously difficult (NP-hard). Therefore, the success of even the first step (i.e. coming up with the target mode) will depend on the form of the function. One can sometimes get around this by choosing to define the performance function in a way that mimics the desired behavior, but will also have some nice features such as convexity (and thus solvable by numerical optimization methods). One such example can be found in correlating Fourier components within the light cone with the loss in cavity Q factor [64, 9]. This restricts the types of performance criteria over which we can optimize.

A second problem is that there may not be a unique optimum to our function. In our 3:1 waveguide splitter example, we arbitrarily chose a parameter for the width of the transverse Gaussian profile because it really did not matter. From a design perspective, as long as the split was 3:1, we are somewhat ambivalent about what the actual width needs to be, as long as the mode remains confined. As such, there would be many such field distributions that are ‘optimum’ for the design goal. However, the inverse problem approach demands the full specification of a single field. Which one of these should we choose? If we can always find a dielectric that produces our desired eigenmode, then this is not a problem. In that case, it simply means that there are multiple designs that are all suitable. Unfortunately, if this is not the case (and so far, it appears that this is the norm), then we have the problem of non-zero residual norm. The question becomes which, if any, of the other field distributions with different width profiles might have been valid, and thus form a solvable inverse problem. There is no way of knowing unless we try them all, which starts to look like trial and error and thus defeat the purpose of the approach.

7.3.1 Valid eigenmode landscape

The critical question to ask becomes how prevalent are these valid eigenmodes? Are we more likely to come up with valid modes or invalid ones that do not readily lead to a solution? If most modes one can design are in fact valid, or at least *approximately* valid, then that is not likely to be a problem. Unfortunately, we can make some strong arguments that the invalid regions are much more prevalent.

Consider a linear operator L such that $Lx = \lambda x$. For small perturbations ΔL, we know that the perturbation to the corresponding eigenvalues and eigenvectors are bounded [78]. Consider the set of all perturbations having some norm $|\Delta L| \leq \xi$, and let the neighborhood of points (corresponding to normalized vectors) on the unit hypersphere that bounds the rotation of a given eigenvector be denoted $\mathcal{S}$. If we treat this purely as a mathematical inverse eigenvalue problem, then we can access any new eigenvector in $\mathcal{S}$ while bounding only the norm of ΔL. Any vector in the neighborhood of an existing eigenvector is a valid eigenvector of some perturbed operator. Indeed, it may seem strange (having framed our design problem in the linear algebra language) to hear about vectors in $\mathcal{C}^N$ that are unfeasible eigenvectors. However, we do not have a purely mathematical inverse problem.

Consider again the Helmholtz operator

$$\Theta^{\eta}_{\mathbf{k},\mathbf{k}'} = \eta_{\mathbf{k}-\mathbf{k}'} \mathbf{k} \cdot \mathbf{k}'. \tag{7.8}$$

In contrast to our general operator L, the Helmholtz operator has $N \times N$ elements and is parameterized by a vector of length N, so we have fewer degrees of freedom to accommodate arbitrary changes to eigenvectors. The structure of the operator means that only a few of these ΔL-type perturbations can take on the valid form of $\Delta\Theta = \delta\eta_{\mathbf{k}-\mathbf{k}'} \mathbf{k} \cdot \mathbf{k}'$, and even fewer of these have an expansion of $\eta(r)$ that can take on physically realizable values. Now, it is certainly true that there exist perturbations to $\eta(r)$ that keep you within $\mathcal{S}$. In fact, the observation of robustness of devices to

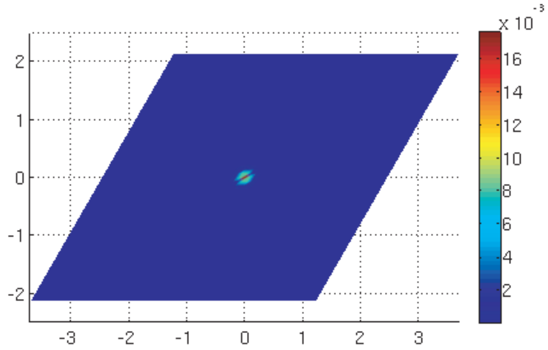

Figure 7.7: Perturbation introduced to $\eta(r)$.

fabrication uncertainty [65] is a good example. However, existence of these modes does not reveal the density of these modes. Our claim is that, due to the rigid form of the Helmholtz operator, the landscape of $\mathbf{H}(\mathbf{r})$ consists mostly of invalid eigenmodes. Therefore, the density of valid modes in $\mathcal{S}$ is small. While we cannot prove this rigorously, we have performed numerical simulations to test our theory.

First, we apply a small perturbation to our canonical h_1 geometry by nominally increasing the radius of the central defect from 0 to 0.03. This corresponds to $|\Delta\eta_k/\eta_k| = 8.6 \times 10^{-4}$ (see figure 7.7). We then solve the forward problem and find the resulting perturbation to the defect mode $|\Delta h_k| = 0.013$ (see figure 7.8). Using the original h_1 eigenvector, we now add to it a perturbation of a much smaller magnitude ($|\Delta a_k| = 10^{-3}$), and proceed to form the inverse problem using this perturbed mode. We minimize the residual norm using the convex optimization scheme, so a non-zero residual norm indicates definitively that the target mode is not a valid eigenmode. In figure 7.9, we plot the distribution of the residual norm for 10000 of these tests. None of these perturbed target 'eigenvectors' corresponded to physically realizable operators. Therefore, the neighborhood surrounding a valid eigenmode appear to be mostly invalid eigenmodes, while the valid modes occupy a set of lower dimensional hypersurfaces within the domain of $\mathbf{H}(\mathbf{r})$'s. As we increase the resolution of the computational grid, the dimensionality increases, further reducing the effective

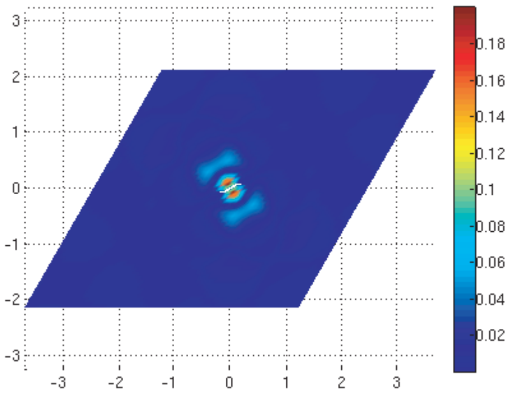

Figure 7.8: Change in eigenmode as a result of perturbation to $\eta(r)$.

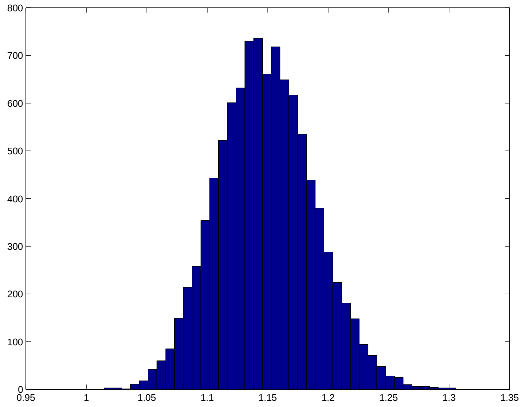

Figure 7.9: Distribution of residual norms of perturbed inverse problems.

density of valid eigenmodes. If we look at figure 7.9 closely, we find the mean value of the residual norm is around 1.15, whereas in section 6.5, the residual norm was closer to 0.25. The number of plane waves used to generate this figure was 3721, whereas we had less than 1000 before, which is another indication that this problem scales badly with the dimension of the state space.

This dramatically changes our concept of the first part of our design process, i.e. the performance optimization of a given mode. The performance metric does not have a continuous domain (set of all normalized vectors, i.e. a hypersphere in $\mathcal{C}^N$), but instead only has little pockets of validity. There is no way of knowing where these pockets are *a priori*, and yet the optimization problem must take this into account. Even when an exact eigenmode is 'close' to our optimized target mode, this distance is not evaluated over the relevant performance metric. It is a general 2-norm distance rather than a weighted distance function that is meaningful to the design goal. We do not know how much of the desired property of the field is lost when we are forced into the valid eigenmode. The limitation imposed by this 'holey' landscape prevents the kind of precision implied by the performance optimization. Given our analysis, it makes the optimization somewhat moot as we are not likely to benefit from it anyway.

We also believe this limitation is quite general despite our bandwidth limitation here. Although computationally impractical, there is no fundamental limitation to the resolution one can use in principle. The argument presented here still holds, and we believe the regions of invalidity dominate more as the dimensionality increases. At infinite resolution, we recover the exact Helmholtz equations, and thus this limitation is clearly independent of the computational model. What we have observed poses a formidable challenge to arbitrary and/or optimal design of PBG structures using an inverse problem method. What is apparent is that this is not a turnkey type design methodology, despite our stated goal of an algorithmic approach to PBG design.

7.4 Conclusion

One obvious improvement that deserves investigation is an alternative to the residual norm as a minimization metric, as discussed throughout this chapter. Since we are dealing with ill-posed problems, we must accept the fact that there is no existence theorem. However, unlike 'standard' inverse problems (where one typically tries to do parameter estimation of the system based on noisy measurements), lack of existence is much more detrimental for our application, whereas lack of uniqueness is not. It does not trouble us that multiple designs all give the same result, whereas the oil prospector cares greatly where the rigs should be located, so a non-unique solution to their inverse problem is much more problematic. When the target mode we seek is not an exact eigenmode, we would like a technique to find a dielectric that supports an eigenmode that is closest to the target mode. As it stands, the entire spectrum contributes to the residual norm, and in a way that is non-optimal. Unfortunately, we have not found a more suitable metric that we can efficiently compute at this point. Until a better metric can be found that decouples the target from the rest of the spectrum, we have elucidated a fundamental limitation to arbitrary and optimal design of photonic structures using an inverse problem based method. Our insight although illustrated through a specific and limited model, turns out to be quite general. Our analysis reveals that the underlying physics fundamentally forbid certain modes from being physically realizable, regardless of human ingenuity. The problem of finding an optimal structure for an arbitrary application, even confirming its optimality, remains an outstanding question in the field of photonic design. Clearly, it is critical to check the validity and quantify the performance of designs obtained through inverse problem methods. We have also established that this general approach is not a turn-key method, but is potentially a very useful tool in instances where one has absolutely no intuition as to how to meet a particular design goal. While the results may not be optimal, they can spawn new geometries that serve as a starting point for other design methods for further fine tuning.

7.4.1 Comparing Tikhonov and Convex Optimization Regularization

Given the inadequacy of the residual norm, the advantage of the COR is not fully realized. We had hoped that non-existence of the solution could be overcome by finding the ‘next best’ solution that does exist. Currently, all COR does for you is tell you that the designed mode is not realizable based on the size of the residual norm. The solution it ends up giving you is not necessarily better or worse than the Tikhonov scheme. As we saw with the noisy h_1 problem, the original solution had a relatively large residual norm, and represents a better solution than some with smaller residual norms. The increased computational efforts make COR less attractive in some circumstances. However, COR is able to handle more sophisticated constraints, whereas Tikhonov cannot, so sometimes we may not have a choice. In addition, COR is less subjective than Tikhonov. Part of the Tikhonov procedure is looking at solutions at various regularization parameters and determining which one is 'best.' We already showed that the L-curve is an objective but unreliable method for finding the best regularization parameter. So that part of the scheme can seem fairly subjective. In addition, the Tikhonov solution will return non-physically realizable values, so the output will need to be fixed. There is some 'slop' inherent in the process, which we may frown upon somewhat, although the residual norm limitations show us that the entire design process is necessarily less rigorous than we had hoped, so we should not be as concerned about it. The COR on the other hand can be fully automated, because the output is guaranteed to be physically realizable. This is especially important with iterative schemes that require many iterations. Both methods will yield good but non-optimal designs, depending on how applicable the residual norm is. The better a metric the residual norm is, the greater the advantage with the COR. This is the reason why we will take finer steps with the COR iterative scheme than with the Tikhonov iterative scheme in dealing with our cQED design problems.