An ADMM Framework for Constrained Bayesian Optimization

Setareh Ariafar
Northeastern University
sariafar@ece.neu.edu

Jaume Coll-Font
Northeastern University
jcollfont@gmail.com

Dana Brooks
Northeastern University
brooks@ece.neu.edu

Jennifer Dy
Northeastern University
jdy@ece.neu.edu

Abstract

Bayesian Optimization (BO) is a powerful machine learning solution for unconstrained optimization when the objective function requires expensive black box evaluation. BO provides a smart sampling procedure that can dramatically reduce the number of required evaluations. Adapting BO to constrained optimization problems where the constraint function is also an expensive black box is attractive, but not a straightforward extension of existing methods. We present a novel approach that leverages natural advantages from a numerical optimization method, the Alternating Direction Method of Multipliers (ADMM). We compare our approach, which we call ADMMBO, against several existing constrained BO methods on two benchmark problems. We reach optimal feasible solutions more rapidly than existing methods and are more robust to starting points.

1 INTRODUCTION

In this paper, we present an approach to solve constrained optimization problems for which the objective function and the constraints are unknown black-box functions that are expensive to evaluate. Our method is set in the context of *Bayesian Optimization* (BO), which is a class of methods solving unconstrained optimization problems while minimizing the number of evaluations (1; 2). BO appropriately selects a sequence of samples to evaluate with an *acquisition function* (AF) that estimates the benefit that would be obtained by sampling at each point in the search space. At every iteration, the "recommended" new point is the maximizer of the acquisition function, which is known and can be optimized with standard optimization techniques. Despite being a relatively new field, several acquisition functions have been proposed. One example is Expected Improvement (EI) (3), which calculates the expectation of improving the best observed objective value so far when sampling any new point. Another, Predictive Entropy Search (PES) (4), computes the expected reduction in the differential entropy of the predictive distribution given a function evaluation at a new point.

Most work in Bayesian Optimization has focused on unconstrained optimization problems. Extending Bayesian Optimization to solve problems with unknown (black box) constraints as well has seen recent interest, although it remains largely unresolved due to its many challenges. For one, there in general is no prior knowledge about the feasible region nor whether the optimal solution, x^*, is likely to be located on its boundary or in the interior. In addition, there are no guarantees that any initial guess is feasible, and in cases where the feasible set is small a considerable number of samples may be evaluated even before starting the full algorithm's first feasible iteration.

Related Work. Constrained Bayesian Optimization approaches can be divided into two main groups: methods that modify the acquisition function and methods based on classical numerical optimization. The first group modifies an unconstrained BO acquisition function so that it simultaneously takes

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

into account the feasibility of a proposed solution along with the objective value. Within this group, Gardner *et.al.* proposed Expected Improvement with Constraints (EIC), which modifies the EI AF by weighting it with the probability of feasibility (5; 6). This method has the advantage of a closed-form expressions for GP models, but high probability of feasibility tends to favor solutions in the interior of the feasible region. This limitation can be problematic when the global optimum is at/near the boundary of the feasible set. A different approach (7) extends PES to formulate a new AF called Predictive Entropy Search with Constraints (PESC) (7; 8). This AF is an expectation, with respect to the posterior distribution of the objective function and constraints evaluated at a candidate point, of the reduction in the differential entropy of the posterior at that point. The advantage of PESC is that it decouples the objective function and constraint terms, allowing them to be optimized separately. Moreover, due to this decoupling, PESC can start from an infeasible point. However, PESC lacks a closed-form solution and thus it requires Expectation Propagation approximation (9), which is relatively hard to implement and might lead to numerical instabilities (7). The second class of methods take advantage of classical optimization techniques based on an unconstrained transformation of the original problem that they can be solved with regular BO methods. To the best of our knowledge, the only existing method following this direction is Augmented Lagrangian BO (ALBO) (10). It solves the Augmented Lagrangian form of the problem, which is unconstrained, and sequentially minimizes this function over x and evaluates the resulting constraint residual to update the Lagrange multipliers. Here, the minimization step is done using Expected Improvement, which yields closed-form solutions after properly reformulating the problem with slack variables (11).

Contributions. Our contribution resides in the second group. We leverage a numerical optimization framework called Alternating Direction Method of Multipliers (ADMM) (12) to formulate a new method, called ADMMBO. In ADMM, we reformulate the problem to include the constraint as an additive term in the objective function that penalizes the objective with $+\infty$ when the solution is infeasible and remains zero otherwise. To solve this now unapproachable optimization problem, ADMM uses variable splitting to decouple the problem into two solvable problems: one that minimizes the objective over the primary variable and a second that minimizes feasibility penalization over an auxiliary variable. The indicator function in the ADMMBO evaluates any feasible point equally, whether located on constraint boundaries or in the interior. Furthermore, ADMMBO sub-problems can flexibly adopt any Bayesian Optimization method of choice. Here we used EI and leveraged closed-form solutions. An additional attractive property of ADMMBO is that we can utilize ADMM optimality conditions as a stopping criterion. ADMMBO shares this property with ALBO, but in the original paper and in our own experiments ALBO usually requires prohibitive iterative computation, hitting a pre-set computational upper limit (budget) before it reaches convergence.

2 ADMMBO Formulation

We are interested in solving the following constrained optimization problem:

$$\begin{aligned} \min_{x \in B} \quad & f(x) \\ \text{s.t.} \quad & c(x) \leq 0, \end{aligned} \tag{1}$$

where $f : R^d \to R$ and $c : R^d \to R$ are unknown Lipschitz continuous functions, and $B \subset R^d$ is a known hyperrectangle. We further assume that we have observed a limited number of function values $F = \{f(x_1), \ldots, f(x_t)\}$ and $C = \{c(x_1), \ldots, c(x_m)\}$. Thus, our goal is to determine a sampling procedure of $f(\cdot)$ and $c(\cdot)$ that sequentially approaches the solution of (1) with the minimum amount of queries possible. To solve problem (1), we reformulate it as an unconstrained optimization $\min_{x \in B} f(x) + \mathbb{1}_\infty\{c(x) > 0\}$ where $\mathbb{1}_\infty\{A\}$ is an indicator function that is ∞ when A is true and zero otherwise. Thus, the constraint $c(\cdot)$ is enforced through a penalty term. Since having the infinity penalty is not computationally tractable, we approximate the objective with a large positive penalty $M < \infty$ and replace $\mathbb{1}_\infty\{c(x) > 0\}$ by $M\mathbb{1}\{c(x) > 0\}$, where $\mathbb{1}\{A\}$ is a binary indicator function that is one when A is true and zero otherwise. The advantage of using this formulation is that we can now solve the problem using ADMM's steps (review of ADMM in the Supplementary Material):

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_{x \in B} \quad f(x) + \frac{\rho}{2}\|x - z^k + \frac{y^k}{\rho}\|_2^2, \\ z^{k+1} &= \operatorname*{argmin}_{z \in B} \quad M\mathbb{1}\{c(z) > 0\} + \frac{\rho}{2}\|x^{k+1} - z + \frac{y^k}{\rho}\|_2^2 \\ y^{k+1} &= y^k + \rho(x^{k+1} - z^{k+1}). \end{aligned} \tag{2}$$

Here, the x update —the *Optimality step*— minimizes the unconstrained objective function with a penalty term that enforces solutions to be close to the feasible region. On the other hand, the z update —the *Feasibility step*— looks for the point in the feasible region that is closest to the unconstrained optimum found in the Optimality step. Since both of the Optimality and Feasibility sub-problems include unknown terms (*i.e.* f-related and c-related terms) we solve them using any unconstrained Bayesian Optimization technique. In the following we will describe an implementation of the ADMMBO framework using Gaussian Processes to model the unknown functions and Expected Improvement yielding to closed-form solutions for both sub-problems. However, the ADMMBO framework can be applied with any combination of probabilistic models and acquisition functions.

2.1 Expected Improvement for the Optimality Step

We call the objective function in the first optimization of (2) as $l(x)$. We model $f(x)$ as a GP, $\tilde{f}(\cdot) \sim GP(0, K(\cdot,\cdot))$. Moreover, since the quadratic term in $l(x)$ is deterministic, the probability distribution of the objective is also a GP, $\tilde{l}(\cdot)$, with mean $\mu(x) = \frac{\rho}{2}\|x - z^k + \frac{y^k}{\rho}\|_2^2$ and the same kernel function as $\tilde{f}(\cdot)$. Thus, the posterior probability of a new function evaluation $l(x_{t+1})$ is a Gaussian distribution with mean $m_l(x_{t+1})$ and variance $\sigma_l(x_{t+1})$, which are determined from the observed samples $L = \{l(x_1), \ldots, l(x_t)\}$ with the standard formulation of GP (13). With this formulation, the expected improvement results in the following closed-form expression:

$$EI(x) = E_{\tilde{l}|L}\big[max(0, l^+ - \tilde{l}(x))\big] = \sigma_l(x)\big(Z\Phi(Z) + \phi(Z)\big), \tag{3}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, $\phi(\cdot)$ is the standard normal probability density function, l^+ denotes the best observed value and $Z = \frac{m_l(x) - l^+}{\sigma_l(x)}$.

2.2 Expected Improvement for the Feasibility Step

We are interested to minimize the objective function in the second optimization of (2) which we call it $h(z)$, given a current set of data, $H = \{h(z_1), \ldots, h(z_m)\}$. To define a probabilistic model for $\tilde{h}(z)$, we assume that $\tilde{c}(\cdot)$ is drawn from a Gaussian Process and model the $I(\tilde{c}(z) > 0)$ as a Bernoulli random variable. The parameter of this variable is then $p[\tilde{c}(z) > 0]$, which is one minus the cumulative distribution function (CDF) of a Gaussian distribution. The quadratic term is a constant value for any z. Thus, $\tilde{h}(z)$ has a shifted Bernoulli distribution described as

$$\tilde{h}(z) = \begin{cases} \frac{\rho}{2M}\|x^{k+1} - z + \frac{y^k}{\rho}\|_2^2 + 1, & \text{with } p\big[\tilde{c}(z) > 0\big] \\ \frac{\rho}{2M}\|x^{k+1} - z + \frac{y^k}{\rho}\|_2^2, & \text{with } p\big[\tilde{c}(z) \leq 0\big] \end{cases} \tag{4}$$

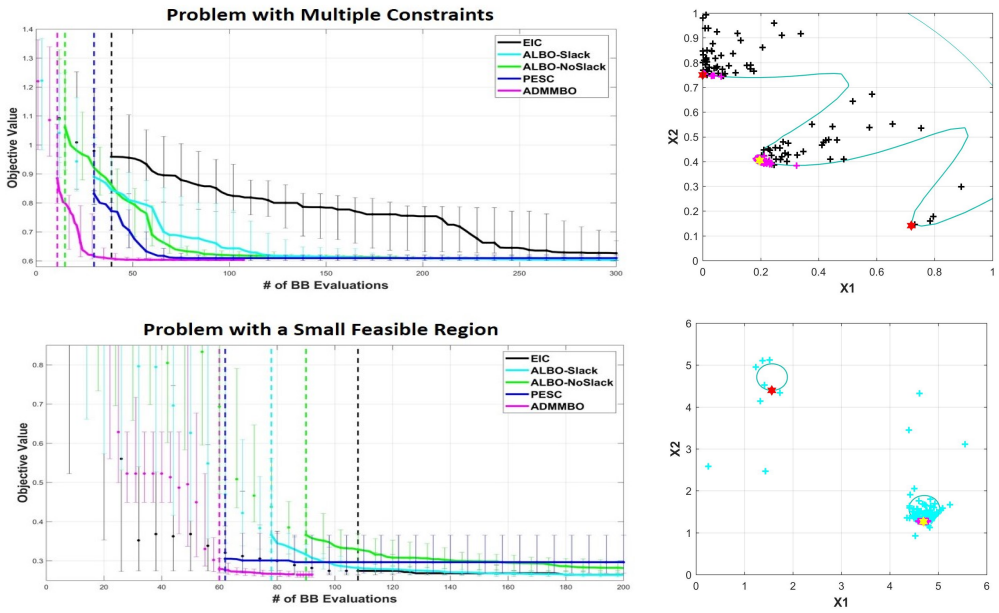
We can now calculate the Expected Improvement for this probabilistic model. Let the minimum best value observed so far be h^+, then EI can be calculated as the expected value of the improvement function (*i.e.* $I(z) = max(0, h^+ - \tilde{h}(z))$) with respect to the shifted Bernoulli distribution:

$$EI(z) = \begin{cases} 0, & \text{if } \; Q(z) \leq 0 \\ I^f(z)p\big[\tilde{c}(z) \leq 0\big], & \text{if } \; 0 < Q(z) \leq 1 \\ I^f(z)p\big[\tilde{c}(z) \leq 0\big] + I^{inf}(z)p\big[\tilde{c}(z) > 0\big], & \text{else} \end{cases} \tag{5}$$

where $Q(z) = h^+ - \frac{\rho}{2M}\|x^{k+1} - z + \frac{y^k}{\rho}\|_2^2$, $I^f(z) = max\big(0, Q(z)\big)$ and $I^{inf}(z) = max\big(0, Q(z) - 1\big)$. Using the Gaussian CDF of $\tilde{c}(.)$, we can exactly and easily compute the Expected Improvement of each point in the feasibility step (see the Supplementary Material for the derivation).

3 EXPERIMENTS & RESULTS

In this section, we highlight four different aspects of the performance of ADMMBO compared to existing methods. First, we are interested in the best feasible objective value obtained for a given budget of black-box (BB) evaluations. Second, we evaluate how rapidly the algorithms approach the optimum as a function of black-box evaluations. Third, we want to report on the number of function evaluations needed to observe the first feasible point, and finally, we are interested in exploring the sensitivity of the algorithms to the (random) initial observations. To report these results, our



figures display curves (in solid lines) of the best objective value $f(x)$ obtained as a function of number of black-box (BB) evaluations. We run our algorithm for 100 random initializations and report the median of the best feasible solution obtained (solid lines on the plots) and also show the 25th and 75th percentiles (to capture variability). Since only feasible solutions are acceptable, we start a curve when the first feasible point is observed. Note that different random initialization runs might need different number of BB evaluations. We show the largest number of BB evaluations (worst case) needed to obtained the first feasible point among all runs, and display this with a vertical dashed line. Hence, the earlier this dashed vertical line occurs on the horizontal axis means better in terms of efficiency in finding a feasible solution. Before starting the median curve, we also report the median and percentiles among all the available runs (the ones that already observed a feasible point) using scattered points to showcase variability of the runs. As we mentioned earlier, we set a maximum over the number of black-box evaluations and run the algorithms until the budget will be exhausted or the optimality conditions, if it has been defined for the algorithm of interest, has been satisfied. We compare ADMMBO against EIC, basic ALBO requiring approximation which we call ALBO-NoSlack, modified ALBO with closed-form solutions which we call ALBO-Slack, and PESC. We chose two benchmark test problems from the Bayesian Optimization literature. **Problem 1** deals with sinusoidal highly non-linear functions as its objective function and constraint resulting in a small feasible region (5). **Problem 2** is a multiple constraint problem optimizing a linear function over a non-linear feasible region based on a quadratic function and a sinusoidal function (11; 10; 7) (details in the supplementary material). Blue solid lines demonstrate the constraint boundaries. Given a fixed budget, magenta crosses denotes the ADMMBO's optima, black crosses show the EIC's optima, and cyan crosses denote ALBO-Slack's, all among 100 runs with random initializations. The yellow star is the real global optimum and the red stars are the real local optima.

Results. Figure 1 shows that among all experiments, ADMMBO is consistently the only method that not only rapidly approaches to the real optimum, it has the capability to stop before exhausting the function evaluation budget. We believe that the reason behind this rapid convergence is because ADMMBO first seeks the unconstrained optimum of the problem, and then looks for the closest point to that optimum which belongs to the feasible set, which turns out to be an effective search strategy. Moreover, ADMMBO also has the least variability. One potential drawback of ADMMBO is initializing the ADMM penalty parameter ρ, the infeasibility penalty coefficient M, the initial Lagrange multiplier y_0, and the initial auxiliary variable z_0. Good initial values will clearly speed up the optimization time. In our examples here we followed the default initialization suggested in (12) and were able to obtain favorable results. However, for more complicated problems, an adaptive initialization policy similar to what ADMM suggests for ρ can make the algorithm less sensitive to the possibility of a poor initialization (12).

References

[1] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of bayesian optimization," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 148–175, 2016.

[2] E. Brochu, V. M. Cora, and N. De Freitas, "A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," *arXiv preprint arXiv:1012.2599*, 2010.

[3] J. Močkus, "On bayesian methods for seeking the extremum," in *Optimization Techniques IFIP Technical Conference*, pp. 400–404, Springer, 1975.

[4] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani, "Predictive entropy search for efficient global optimization of black-box functions," in *Advances in neural information processing systems*, pp. 918–926, 2014.

[5] J. R. Gardner, M. J. Kusner, Z. E. Xu, K. Q. Weinberger, and J. P. Cunningham, "Bayesian optimization with inequality constraints.," in *ICML*, pp. 937–945, 2014.

[6] M. A. Gelbart, J. Snoek, and R. P. Adams, "Bayesian optimization with unknown constraints," *arXiv preprint arXiv:1403.5607*, 2014.

[7] J. M. Hernández-Lobato, M. Gelbart, M. Hoffman, R. Adams, and Z. Ghahramani, "Predictive entropy search for bayesian optimization with unknown constraints," in *International Conference on Machine Learning*, pp. 1699–1707, 2015.

[8] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, pp. 2951–2959, 2012.

[9] T. P. Minka, *A family of algorithms for approximate Bayesian inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[10] R. B. Gramacy, G. A. Gray, S. Le Digabel, H. K. Lee, P. Ranjan, G. Wells, and S. M. Wild, "Modeling an augmented lagrangian for blackbox constrained optimization," *Technometrics*, vol. 58, no. 1, pp. 1–11, 2016.

[11] V. Picheny, R. B. Gramacy, S. Wild, and S. Le Digabel, "Bayesian optimization under mixed constraints with a slack-variable augmented lagrangian," in *Advances in Neural Information Processing Systems*, pp. 1435–1443, 2016.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] C. E. Rasmussen and C. K. Williams, *Gaussian processes for machine learning*, vol. 1. MIT press Cambridge, 2006.

Supplementary Material

ADMM REVIEW

ADMM solves problems in the form

$$\min_{x \in B} \quad f(x) + g(x), \tag{6}$$

where $x \in R^d$, $f, g : R^d \to R$ and, without loss of generality, $B \subset R^d$ is a known bounded hyperrectangle. In order to solve this optimization problem, ADMM introduces an auxiliary variable $z \in R^d$ and reformulates the problem as:

$$\begin{aligned} \min_{x,z \in B} \quad & f(x) + g(z) \\ \text{s.t.} \quad & x = z. \end{aligned} \tag{7}$$

Since the equality constraint enforces x and z to be the same, the problem (7) remains equivalent to the problem (6). The Augmented Lagrangian function for problem (7) is:

$$L_\rho(x, z, y) = f(x) + g(z) + y^T(x - z) + \frac{\rho}{2}\|x - z\|_2^2. \tag{8}$$

The advantage of this new formulation is that (8) is separable with respect to $f(.)$ and $g(.)$ and can be solved by alternatively optimizing over the following sub-problems:

$$\begin{aligned} x^{k+1} &= \operatorname*{argmin}_x \quad L_\rho(x, z^k, y^k), \\ z^{k+1} &= \operatorname*{argmin}_z \quad L_\rho(x^{k+1}, z, y^k), \\ y^{k+1} &= y^k + \rho(x^{k+1} - z^{k+1}), \end{aligned} \tag{9}$$

where ρ is a positive real parameter. For a detailed review see (12).

We want to emphasize an important distinction between this formulation and ALBO. Both use an Augmented Lagrangian function. However ALBO computes the Augmented Lagrangian function over the base problem with the unknown constraints included resulting in a prohibitive penalty term, $\max(0, c(x)^2)$, in the Augmented Lagrangian function. However, as can be seen above, ADMMBO forms the Augmented Lagrangian function over an equivalent problem with a deterministic constraint over the primary and auxiliary variables, resulting in a deterministic penalty term, $\frac{\rho}{2}\|x - z\|_2^2$, in the Augmented Lagrangian function. Thus , in the context of BO with unknown objective and constraint, this means that ALBO needs to treat the constraint as a stochastic problem while ADMMBO treats the constraint as deterministic with a simple quadratic penalty.

Optimality Conditions

Primal feasibility and dual feasibility are the necessary and sufficient conditions for an optimal ADMM solution. To evaluate these feasibility conditions ADMM defines the primal and dual residuals

$$\begin{aligned} r^{k+1} &= x^{k+1} - z^{k+1}, \\ s^{k+1} &= \rho(z^{k+1} - z^k), \end{aligned} \tag{10}$$

respectively. The algorithm is considered to have converged when the residuals are sufficiently small:

$$\begin{aligned} \|r^{k+1}\|_2 &\leq \epsilon^{primal}, \\ \|s^{k+1}\|_2 &\leq \epsilon^{dual}, \end{aligned} \tag{11}$$

where ϵ^{primal} and ϵ^{dual} are pre-defined optimality tolerances. Here we choose them following a heuristic suggested in (12). ADMM iterates through the sub-problems in (9) until it either satisfies both stopping criteria in (11) or reaches a preset maximum number of iterations. In practice there are heuristic approaches to update the penalty parameter ρ across iterations (12).

Deriving the Expected Improvement for the Feasibility Step

$$
\begin{aligned}
&EI(z) = \underset{\tilde{h}|H}{E}(max(0, h^+ - \tilde{h}(z))) = \\
&max\left(0, h^+ - \frac{\rho}{2M}\|x^{k+1} - z + \frac{y^k}{\rho}\|_2^2 - 1\right) p\big[\tilde{c}(z) > 0\big] \\
&+ max\left(0, h^+ - \frac{\rho}{2M}\|x^{k+1} - z + \frac{y^k}{\rho}\|_2^2\right) p\big[\tilde{c}(z) \leq 0\big].
\end{aligned} \tag{12}
$$

The value of this equation depends on the range of the distance between the best value observed so far and the scaled shifted distance between z and the optimum found by the optimality step x^{k+1} (*i.e.* $h^+ - \frac{\rho}{2M}\|x^{k+1} - z + \frac{y^k}{\rho}\|_2^2$) which we denote as $Q(z)$. If the mentioned term is non-positive, both improvements and their corresponding expectations are zero. If the $Q(z)$ lies between zero and one, the first improvement is zero while the second one has a positive value. Otherwise when this term is larger than one, both improvements are positive. Thus, we can simplify equation (12) as (5).

Mathematical Definition of Test Problems

In the following we will mathematically explain the benchmark problems we tested in this paper. Totally there are two objective functions, one linear and one non-linear, with four constraint functions.

$$f_1(x) = \sum_{i=1}^{d} sin(x_1) + x_2 \tag{13}$$

$$f_2(x) = \sum_{i=1}^{d} x_i \tag{14}$$

$$c_1(x) = sin(x_1)sin(x_2) + 0.95 \tag{15}$$

$$c_2(x) = -x_1^2 - x_2^2 + 1.5 \tag{16}$$

$$c_3(x) = \frac{1}{2}sin(2\pi(x_1^2 - 2x_2)) + x_1 + 2x_2 + 1.5 \tag{17}$$

Problem 1 optimizes the $f_1(x)$ over $c_1(x)$ where $x_i \in [0\ 6], i = 1, \ldots, d$. **Problem 2** optimizes the $f_2(x)$ over two constraints, i.e. $c_2(x)$ and $c_3(x)$. $x_i \in [0\ 1], i = 1, \ldots, d$. We used a squared exponential GP kernel, i.e. $K(\cdot, \cdot)$ for our test problems and set the hyperparameters as suggested by the paper which visited that problem first. For example for Problem 2, we used $\delta_1 = \delta_2 = 0.025$ following (7). Among all experiments, we started from two initial points for both the objective function and constraints.