A Linear Programming Approach to Computational Bayesian Persuasion

Anthony Lee Zhang

June 29, 2017

1 Summary

Consider a Bayesian persuasion problem in which the state of nature ω is one-dimensional, the receiver takes an action which is a function of the posterior mean of the state of nature, and the sender has a state-independent utility function over the receiver's action. The sender's problem is to choose an optimal distribution of posterior means. A characterization theorem from Gentzkow and Kamenica [2016] shows that the sender's optimization problem is linear in the posterior mean distribution G, and is subject to linear constraints on G. I show that the sender's optimization problem can be approximated by a tractable linear program, and I use this to numerically solve a number of examples of such Bayesian persuasion problems.

To my knowledge, the idea for applying linear programming techniques to solve optimization problems with stochastic dominance constraints originates in Dentcheva and Ruszczynski [2003], as a method for optimizing asset portfolios, subject to the constraint that they are preferred to some benchmark asset in the sense of stochastic dominance. Recently, a number of other authors, including Dworczak and Martini [2016] and Kolotilin [2016], have analyzed similar linear programming formulations of Bayesian persuasion problems. Relative to these papers, to my knowledge, my grid approximation results of Theorem 2 are novel, though fairly elementary.

2 Model

The model is identical to that of Gentzkow and Kamenica [2016]. The state of nature is a random variable $\omega \in [0, 1]$. There are two agents, sender and receiver, with common prior F. Sender observes ω and chooses a *signal structure* π, which is a random variable jointly distributed with ω. Receiver does not observe ω, but observes the joint distribution of π and ω, as well as the realization of π. Receiver takes an action α which is a function of the posterior mean/conditional expectation of the state of nature $E[\omega \mid \pi]$, which induces some state-independent utility for sender $v(\alpha(E[\omega \mid \pi]))$. Hence, without loss of generality, we can abstract away from the particular signals and actions, and think of the sender as choosing a distribution G of posterior means $\gamma = E[\omega \mid \pi]$, which she has some utility function for: $u(\gamma) = v(\alpha(\gamma))$. Sender's optimization problem is then to choose G to maximize:

$$\int u(\gamma)\ dG(\gamma)$$

subject to the constraint that G can be induced as the marginal distribution of the posterior mean under some signal structure.

3 Characterization of feasible posterior mean distributions

In order to characterize the set of feasible posterior mean distributions G, we have to answer the following question: given marginal distributions F and G, under what circumstances can we construct random variables X, Y jointly distributed on some probability space $(\Omega, \mathscr{F}, \mu)$ such that, marginally, $X \sim F$, $Y \sim G$, and $Y = E(X \mid Y)$? As Gentzkow and Kamenica [2016] show, following among others Strassen [1965], there is a fairly simple characterization theorem for this set. I summarize a number of equivalent statements of this characterization in the following Theorem:

Theorem 1. *Given marginal distributions F and G, one can construct random variables X, Y jointly distributed on some probability space $(\Omega, \mathscr{F}, \mu)$ such that, marginally, $X \sim F$, $Y \sim G$, and $Y = E(X \mid Y)$, if and only if one of the following equivalent conditions hold:*

1. *Kamenica-Gentzkow formulation:*

$$\int_0^a F(x)\,dx \geqslant \int_0^a G(x)\,dx \;\forall a, \tag{1}$$

$$\int_0^1 F(x)\,dx = \int_0^1 G(x)\,dx$$

2. *F and G have the same mean, and for all $a \in [0,1]$:*

$$\int_0^1 (a-x)\,1_{x\leqslant a}\,dF(x) \geqslant \int_0^1 (a-x)\,1_{x\leqslant a}\,dG(x)$$

3. *F and G have the same mean, and for all increasing convex functions $\phi(x)$:*

$$\int_0^1 \phi(x)\,dF(x) \geqslant \int_0^1 \phi(x)\,dG(x)$$

4. *F is a mean-preserving spread of G.*

5. *F second-order stochastically dominates G, and F and G have the same mean.*

Remark. Equivalence between 1. and 2. can be shown by an integration by parts (see Appendix A). 3. immediately implies 2., and 2. implies 3. because any convex function can be approximated arbitrarily closely by linear combinations of the functions $(a-x)\,1_{x\leqslant a}$; see Dentcheva and Ruszczynski [2003], Proposition 2.2, for a proof of this. 4. and 5. are equivalent to the others by definition; 4. can be defined in terms of 1., and 5. is often defined in terms of 3.

4 An LP approximation to optimal Bayesian persuasion

Using formulation 2., the sender's problem can be written as

$$\max_{G(\cdot)} \int u(\gamma)\,dG(\gamma)$$

$$\text{s.t.} \int_0^1 (a-x)\,1_{x\leqslant a}\,dG(x) \leqslant \int_0^1 (a-x)\,1_{x\leqslant a}\,dF(x) \;\; \forall a \in [0,1]$$

In this problem, the choice parameter G is a probability measure, which is infinite dimensional, and there are infinitely many constraints, hence standard numerical optimization techniques don't immediately

apply. However, both the objective and the constraints are linear in the probability measure G, suggesting that a linear programming approximation may be possible.

I propose to approximate this with a problem with a finite number of choice parameters and constraints. Fix some grid width δ; for simplicity suppose $\delta = \frac{1}{n}$ for some n. We will optimize over *grid distributions* G_δ, which place probabilities $p_0, p_1, \dots p_n$ on the $n+1$ points $\{0, \delta, 2\delta \dots n\delta\}$, and 0 probability elsewhere. We will enforce constraints

$$\int_0^1 (a - x)\, 1_{x \leqslant a}\, dG(x) \leqslant \int_0^1 (a - x)\, 1_{x \leqslant a}\, dF(x)$$

also only for $a \in \{0, \delta, 2\delta \dots n\delta\}$. Hence, our optimization problem is:

$$\max_{p_0, \dots p_n} \sum_{i=0}^{n} p_i u_i \text{ s.t. :}$$

$$p_i \geqslant 0 \ \forall i$$

$$\sum_{i=0}^{n} p_i = 1$$

$$\sum_{i=0}^{j} (j\delta - i\delta)\, p_i \leqslant \int_0^1 (j\delta - x)\, 1_{x \leqslant a}\, dF(x) \ \ \forall j \in \{0, 1, \dots n\} \tag{2}$$

$$\sum_{i=0}^{n} i\delta p_i = \int_0^1 x\, dF(x) \tag{3}$$

In the following sections I show that, under weak smoothness constraints on the objective function u (in particular allowing for finitely many discontinuities in u, which accomodates the case in which receiver takes finitely many discrete actions), for any exactly optimal distribution G^*, there is an approximating grid distribution G_δ which achieves approximately the same utility, and approximately satisfies all constraints, with both approximation errors decreasing to 0 linearly with δ. Hence the outcome of the finite optimization problem is "close" to the solution of the original infinite problem.

4.1 Approximation Theorem

The primitives of the problem are the prior distribution F of the signal and the utility function $u : [0, 1] \to \mathbb{R}$. We will require that u is discontinuous at most at a finite number of points $\{a_1 \dots a_m\}$ in the unit interval. On all subintervals for which u is continuous, it is Lipschitz continuous with parameter K. Also, we will assume that the grid width δ is small enough that $\delta < \min\{a_i - a_{i-1}\}$, so at least one grid point falls in every continuity interval of u.

Theorem 2. *Under the assumptions above, there exists some grid distribution G_δ which achieves within $K\delta$ of the optimal value of G^*, and violates all constraints by at most 3δ, that is:*

$$\left| \int_0^1 U(x)\ dG^*(x) - \int_0^1 U(x)\ dG_\delta(x) \right| \leqslant 2K\delta$$

$$\left| \int_0^1 x\, dG^*(x) - \int_0^1 x\, dG_\delta(x) \right| \leqslant 2\delta$$

$$\left| \int_0^1 (x - a)\, 1_{x > a}\, dG^*(x) - \int_0^1 (x - a)\, 1_{x > a}\, dG_\delta(x) \right| \leqslant 3\delta \ \forall a$$

Proof. See Appendix B. □

5 Computational Results

In Figures 1-6 I show the results of the algorithm on a number of utility functions for sender. In all cases the prior F is uniform on $[0, 1]$.

In Figures 1, 2 and 3, the utility function is a step function. Figure 1 displays the basic BP problem where a judge wants to convince receiver that, with above 85% probability the suspect is guilty; in this case, an optimal solution is a 3-point distribution with support $0.85, 0.5, 0.15$. Figure 2 displays the numerical example discussed in Gentzkow and Kamenica [2016]. In Figure 3 we show a "convex-like" step function, in which sender prefers more extreme beliefs, but in a step-function manner. The sender's optimal solution concentrates probability on the step-function cutoffs $0.1, 0.3, 0.7, 0.9$.

Figures 4 and 5 show the standard cases where u is fully concave or convex; the optimal strategy is to reveal no information or all information, respectively. As an example of a partially concave and partially convex function, figure 6 shows the case where u is a sine function. The sender's optimal strategy is, roughly speaking, to reveal information in the convex part and not reveal information in the concave part of the distribution. But this doesn't exactly hold – information is revealed only above $x = 0.75$, not the entire convex region, and as a result the mass point in the concave region falls at approx. $x = 0.3$ rather than the peak of the sine. I believe this illustrates that, in general, the solution to this class of problems can be complex, and it seems like in general does not admit analytical solution; hence computation seems like an attractive approach.

In addition, in each plot I show the Lagrange multipliers from the linear program as a function of a. In most cases, only a few Lagrange multipliers are positive, meaning only a few of the convex dominance constraints bind. I believe the Lagrange multipliers can be interpreted roughly as saying that, to first order, the sender would benefit from more "dispersion" in the prior locally around each a for which the Lagrange multipliers are positive, and the magnitude of the Lagrange multiplier represents the magnitude of the sender's gain from local dispersion.

A Equivalence between conditions 1. and 2.

Suppose that:

$$\int_0^a F(x)\,dx \geqslant \int_0^a G(x)\,dx \;\forall a,$$
$$\int_0^1 F(x)\,dx = \int_0^1 G(x)\,dx$$

Use integration by parts:

$$\int_0^a F(x)\,dx = (x-a)\,F(x)\,|_0^a - \int_0^a (x-a)\,dF(x)$$
$$= \int_0^a (a-x)\,dF(x) = \int_0^1 (a-x)\,1_{x\leqslant a}\,dF(x)$$

Hence, this is equivalent to:

$$\int_0^1 (a-x)\,1_{x\leqslant a}\,dF(x) \geqslant \int_0^1 (a-x)\,1_{x\leqslant a}\,dG(x) \;\;\forall a,$$
$$\int_0^1 x\,dF(x) = \int_0^1 x\,dG(x)$$

Figure 1: Two-level utility function

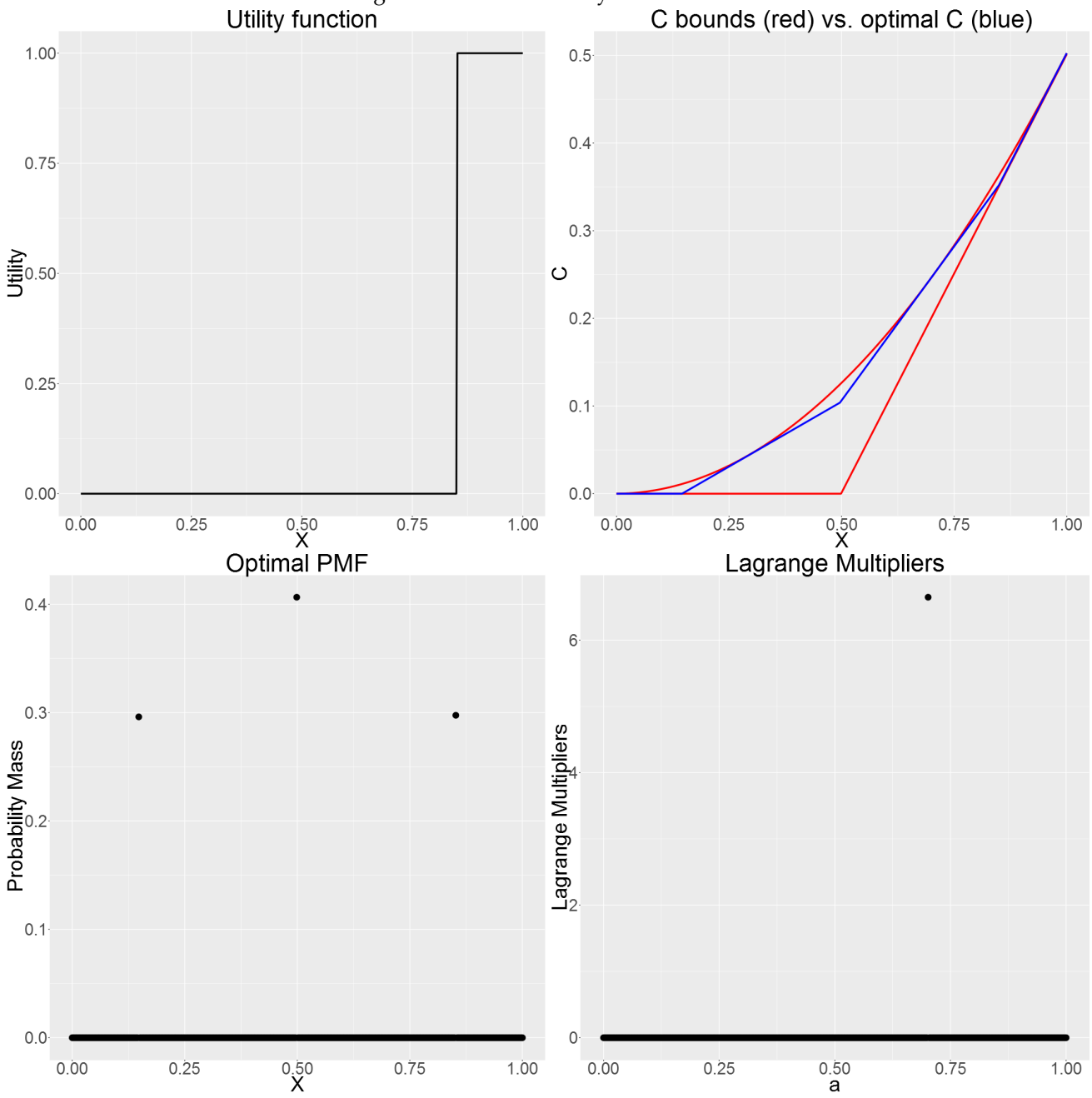


Figure 2: Three-level utility function

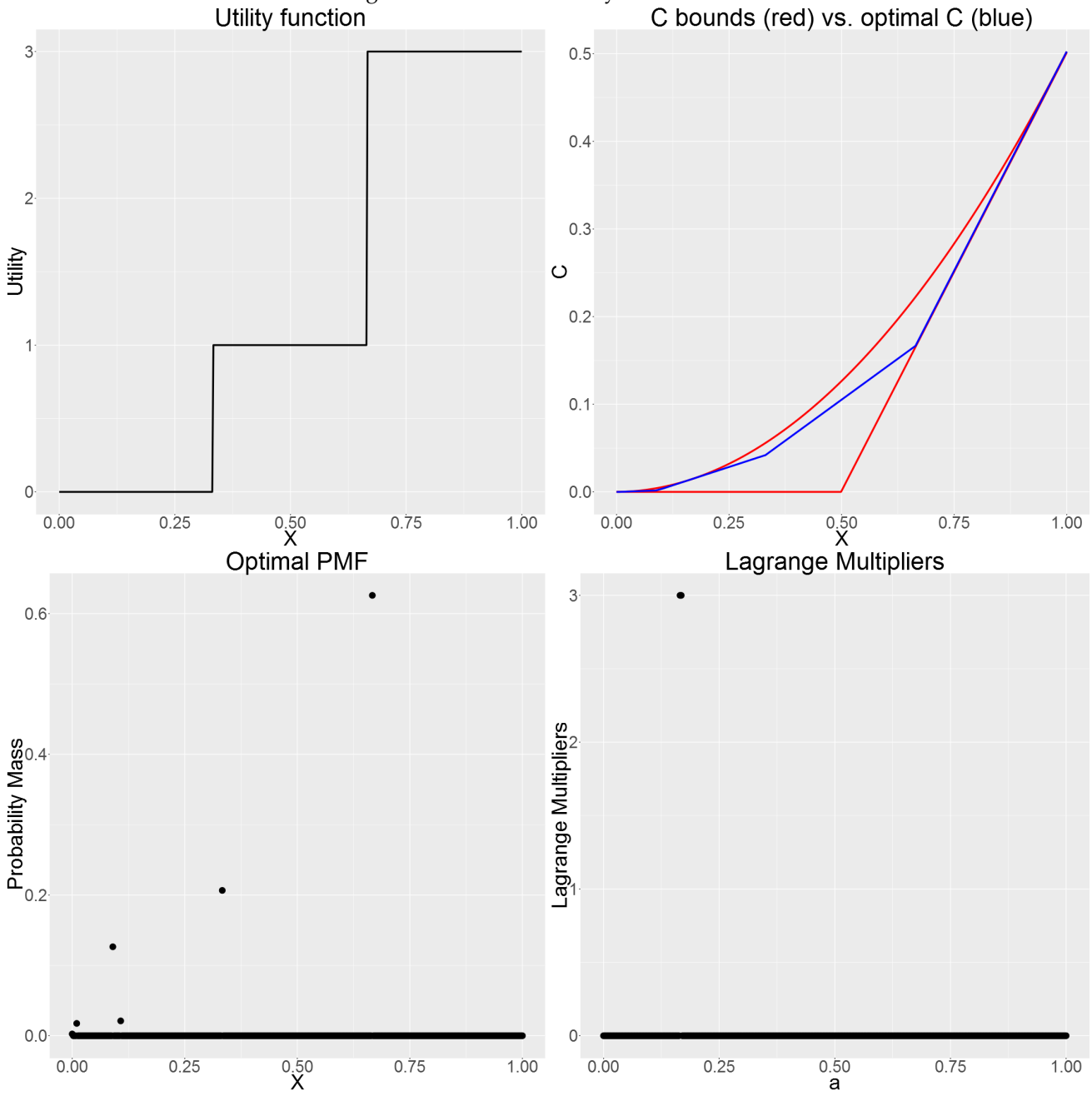


Figure 3: Five-level utility function

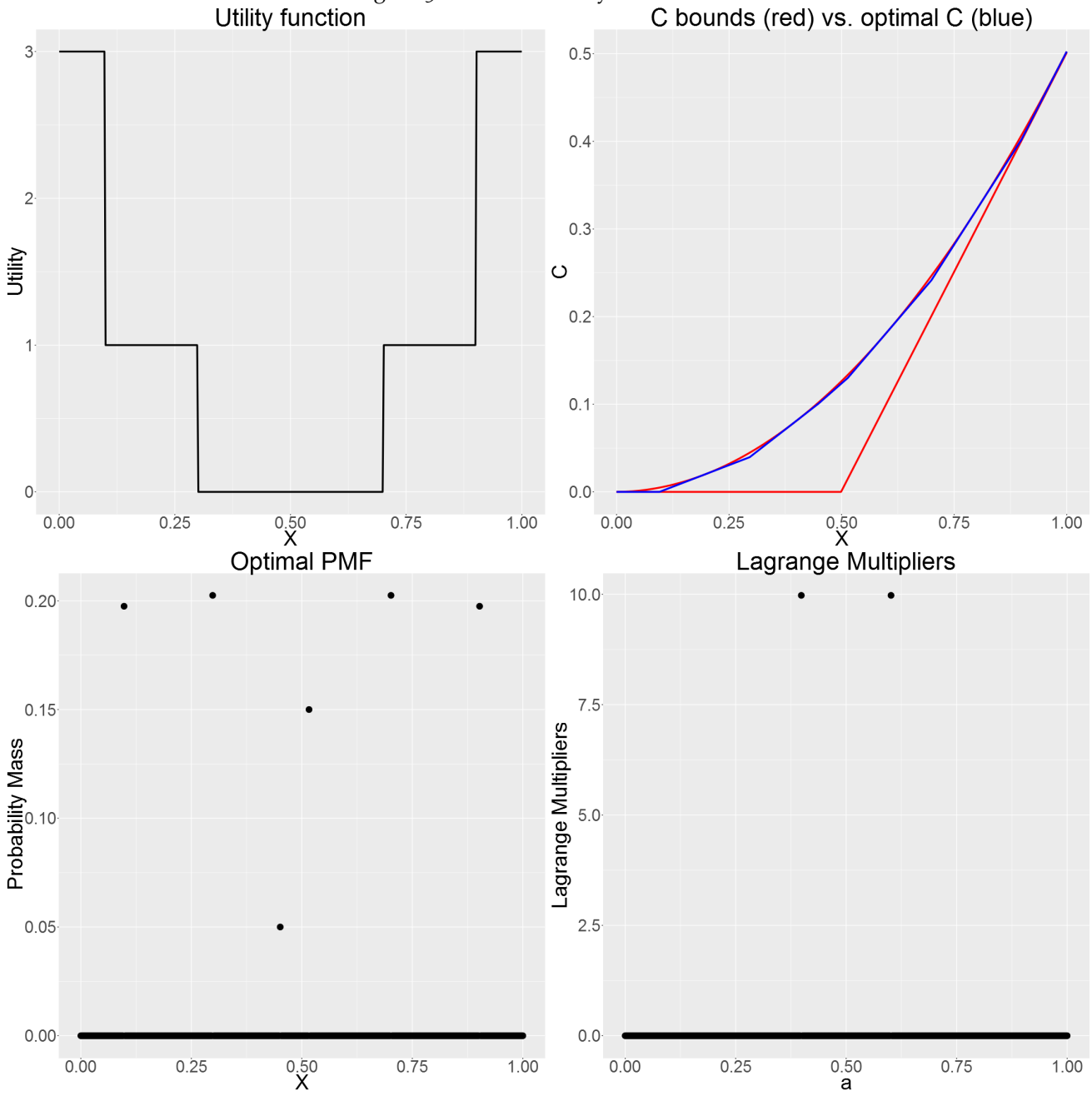


Figure 4: Concave utility function

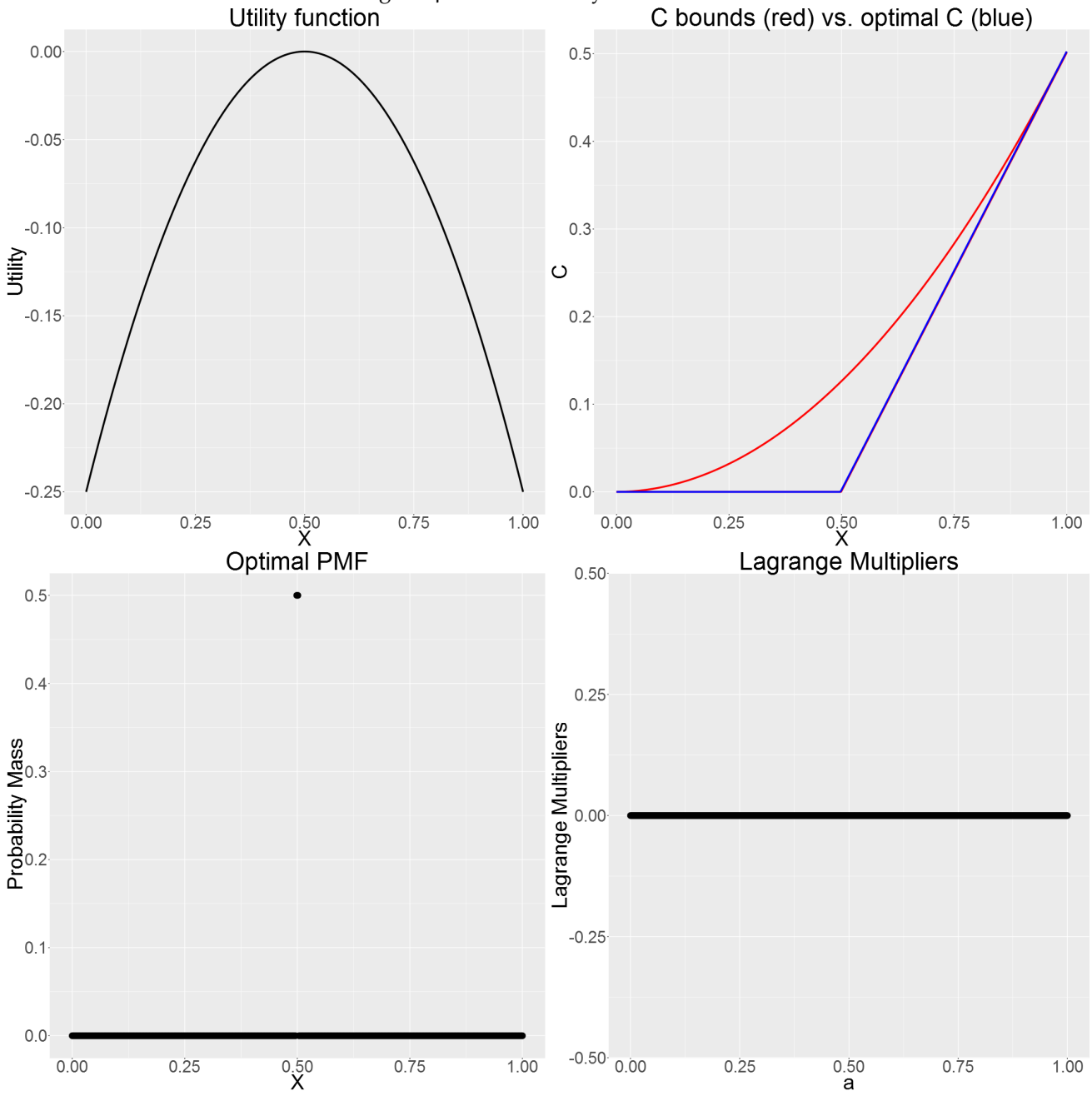


Figure 5: Convex utility function

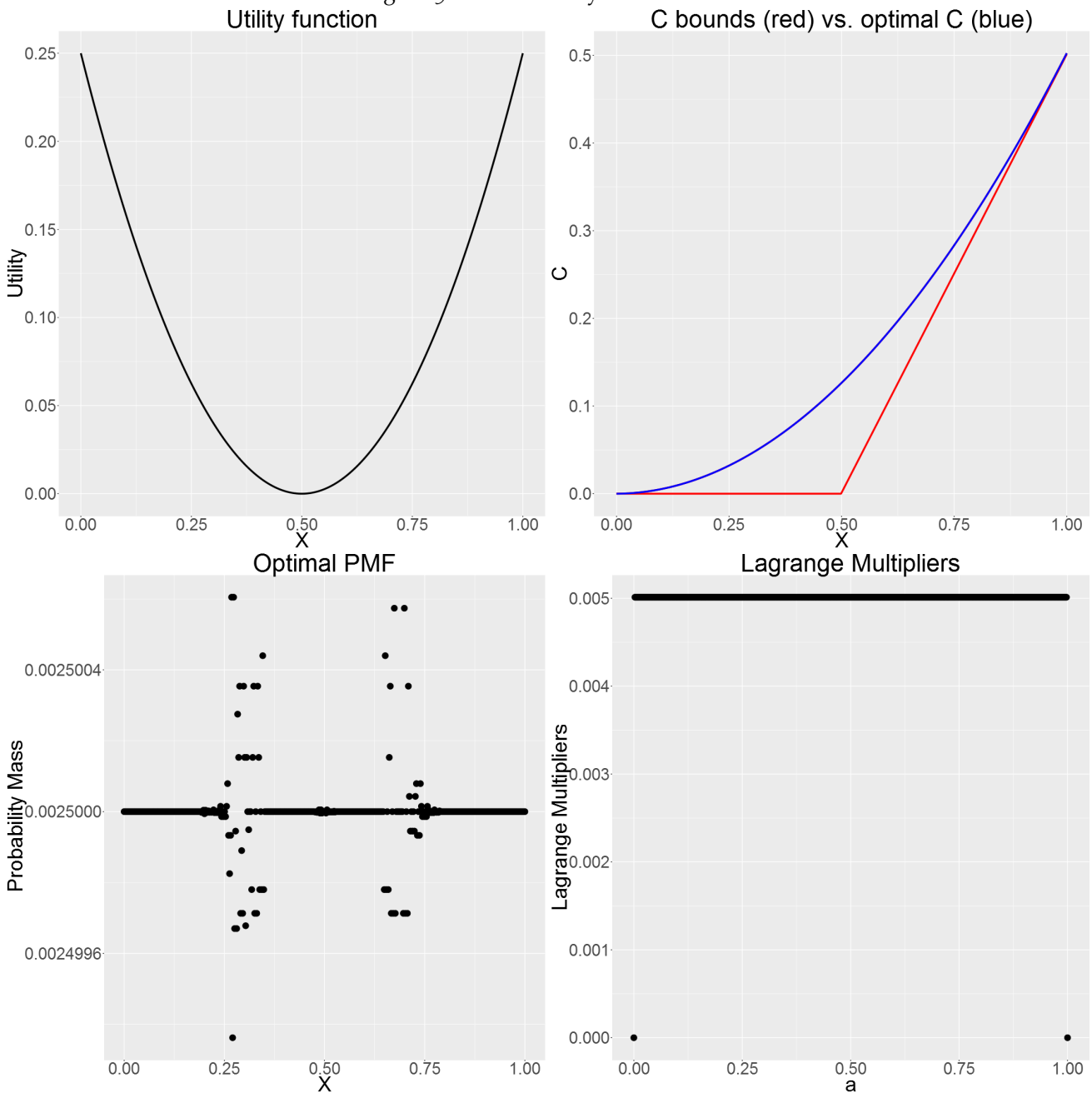
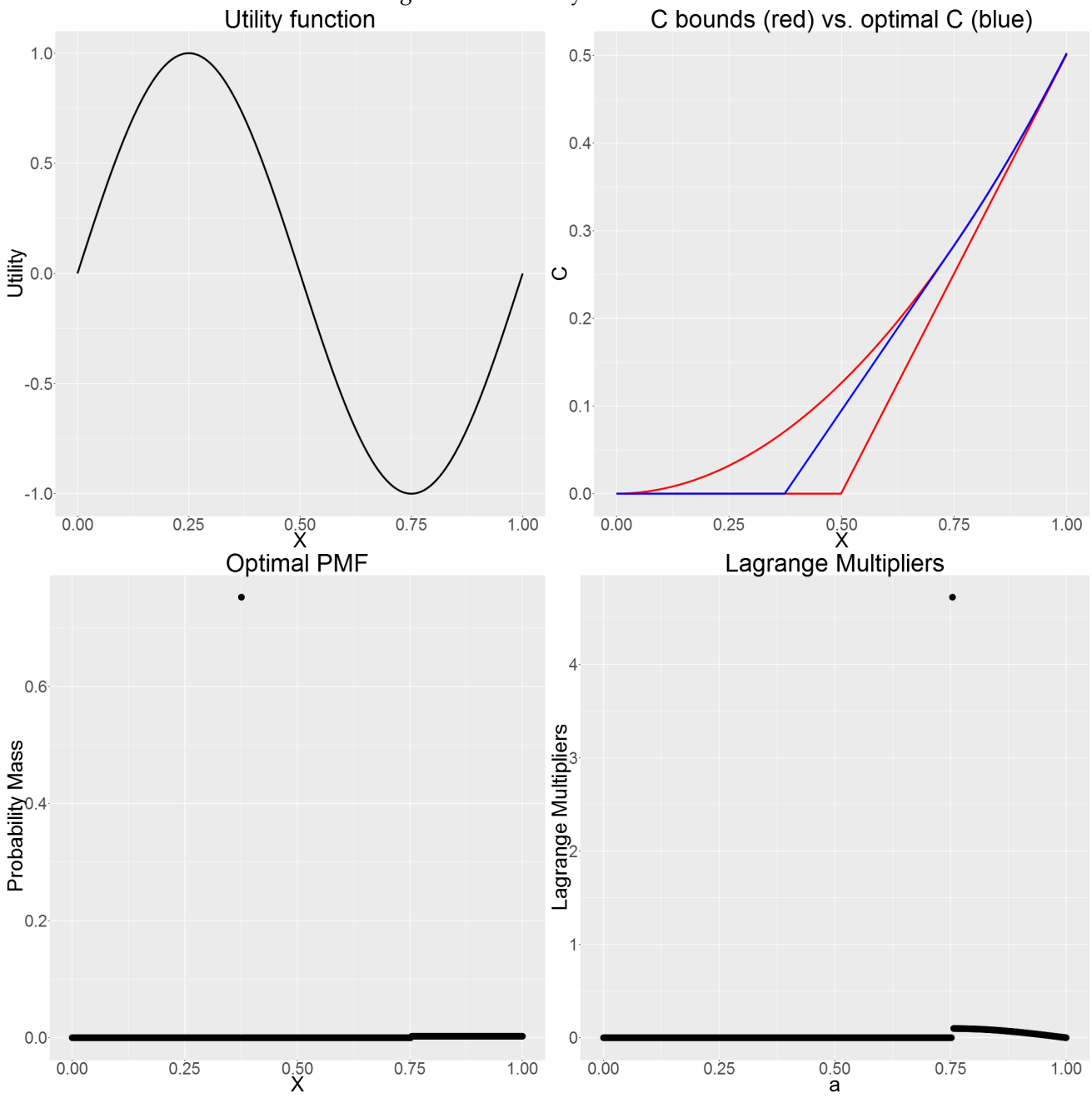


Figure 6: Sine utility function



B Proof of Theorem 2

Define the $m+1$ subintervals $\tilde{I}_1 \dots \tilde{I}_{m+1}$ as $\tilde{I}_1 = (0, a_1)$, $\tilde{I}_2 = (a_1, a_2) \dots \tilde{I}_{m+1} = (a_m, 1)$. The union of the $\tilde{I}_j$ intervals covers the unit interval minus the discontinuity points $\{a_1 \dots a_m\}$. We will attach the discontinuity points to the intervals in a way that the sender prefers. That is to say, suppose that $u(a_1 + \epsilon) > u(a_1 - \epsilon)$. Then we will attach a_1 to the right-side interval $\tilde{I}_2$, defining $I_2 = \tilde{I}_2 \cup \{a_1\}$. In similar fashion, define for each j

$$I_j = \tilde{I}_j \cup \left\{ \left\{ a_j \right\} \text{ if } u\left(a_j - \epsilon\right) > u\left(a_j + \epsilon\right) \right\} \cup \left\{ \left\{ a_{j-1} \right\} \text{ if } u\left(a_{j-1} + \epsilon\right) > u\left(a_{j-1} - \epsilon\right) \right\}$$

Now each of the grid points $x_i = \delta i$, $i \in \{1, 2, \dots n\}$ lives in exactly one of the I_j intervals. Consider the exactly optimal posterior mean distribution G^*. For each interval I_j, we will redistribute the probability mass that falls within I_j to the point masses x_i that fall within I_j, as follows. Suppose that grid points $x_s, x_{s+1}, \dots x_{s+t}$ fall in I_j. Then, for each $x_i \in \{x_s, x_{s+1}, \dots x_{s+t}\}$, we will define an "absorption interval" L_i as follows:

- $L_s = I_j \cap [a_{j-1}, x_s + \frac{\delta}{2})$
- $L_{s+t'} = I_j \cap [x_{s+t'} - \frac{\delta}{2},\ x_{s+t'} + \frac{\delta}{2})\ \forall t' < t$
- $L_{s+t} = I_j \cap \left[x_{s+t} - \frac{\delta}{2}, a_j\right]$

Then, we construct G_δ by assigning to x_i all probability mass that falls in the absorption interval L_i.

In constructing G_δ from G^*, we redistribute probability mass within each interval I_j, such that the points closest to the boundaries of I_j absorb all probability mass near the boundary, and all other points absorb probability mass within a ball of radius $\frac{\delta}{2}$. Hence, probability mass is conserved within each continuity interval, so no probability mass "jumps" across discontinuities. Also, no unit of probability mass moves more than δ in the x-dimension; since the objective u is Lipschitz of parameter K, this translates into a bound for the loss in utility, as well as the differences in constraint values, from G_δ relative to G^*.

Claim. The difference between the expected utility of G^* and G_δ is bounded above by $K\delta$, that is:

$$\left| \int_0^1 U(x)\ dG^*(x) - \int_0^1 U(x)\ dG_\delta(x) \right| \leqslant 2K\delta$$

The differences between the constraint values under G^* and G_δ, that is, $\left|\int_0^1 x\, dG^*(x) - \int_0^1 x\, dG_\delta(x)\right|$ and $\left|\int_0^1 (x-a)\, 1_{x>a}\, dG^*(x) - \int_0^1 (x-a)\, 1_{x>a}\, dG_\delta(x)\right|$, are bounded above by 2δ.

Proof. For notational convenience define $\mu^*(L_i) = \int_{L_i} dG^*(x)$; that is, $\mu^*(L_i)$ is the total probability mass assigned to L_i under G^*. On each L_i, we have:

$$\min_{x \in L_i} U(x)\, \mu_\delta(L_i) \leqslant \int_{L_i} U(x)\ dG_\delta(x) \leqslant \max_{x \in L_i} U(x)\, \mu_\delta(L_i),$$

$$\min_{x \in L_i} U(x)\, \mu^*(L_i) \leqslant \int_{L_i} U(x)\ dG^*(x) \leqslant \max_{x \in L_i} U(x)\, \mu^*(L_i).$$

However, by construction $\mu_\delta(L_i) = \mu^*(L_i)$. Also, by our Lipschitz assumption, since the intervals L_i have length bounded above by 2δ, $\max_{x \in L_i} U(x) - \min_{x \in L_i} U(x) \leqslant 2K\delta$. Hence,

$$\left| \int_{L_i} U(x)\ dG^*(x) - \int_{L_i} U(x)\ dG_\delta(x) \right| \leqslant 2K\delta\mu^*(L_i)$$

Summing over all intervals L_i, and using that $\sum_i \mu^*(L_i) = \int dG^*(x) = 1$, gives the desired result for U. For the constraints, note that the functions x and $(x-a)\, 1_{x>a}$ are Lipschitz with $K = 1$, hence the same argument implies that the constraint values change by at most 2δ. □

However, the approximation scheme cannot guarantee bounded utility loss under full constraint satisfaction; one can create examples in which any grid distribution which exactly satisfies convex dominance constraints has arbitrarily high loss relative to G^*.

In practice, for computational convenience, we will run the optimization using a δ-grid probability distribution as well as constraints enforced only on the δ-grid. Any function which exactly satisfies convex dominance constraints for all $a = \{\delta, 2\delta, 3\delta \ldots\}$ almost satisfies the constraints for all a, as the following claim shows:

Claim. Here, we use formulation 1 of the convex dominance condition. Suppose that the inequalities $\int_0^a F(x)\,dx \geqslant \int_0^a G(x)\,dx$ hold for all $a = \{\delta, 2\delta, 3\delta \ldots\}$. Then, for any $a' \in [0, 1]$, $\int_0^{a'} F(x)\,dx + \delta \geqslant \int_0^{a'} G(x)\,dx$.

Proof. Any a lies in the interval $[m\delta, (m+1)\delta]$ for some m. Then, we have:

$$\int_0^{a'} G(x)\,dx \leqslant \int_0^{(m+1)\delta} G(x)\,dx \leqslant \int_0^{(m+1)\delta} F(x)\,dx = \int_0^{m\delta} F(x)\,dx + \int_{m\delta}^{(m+1)\delta} F(x)\,dx$$

Using that $F(x)$ is bounded above by 1,

$$\leqslant \int_0^{m\delta} F(x)\,dx + \int_{m\delta}^{(m+1)\delta} 1\,dx \leqslant \int_0^{a'} F(x)\,dx + \delta$$

proving the desired result. □

References

D. Dentcheva and A. Ruszczynski. Optimization with Stochastic Dominance Constraints. *SIAM Journal on Optimization*, 14(2):548–566, January 2003. ISSN 1052-6234. doi: 10.1137/S1052623402420528. URL http://epubs.siam.org/doi/abs/10.1137/S1052623402420528.

Piotr Dworczak and Giorgio Martini. A Duality Approach to Bayesian Persuasion. SSRN Scholarly Paper ID 2785970, Social Science Research Network, Rochester, NY, May 2016. URL https://papers.ssrn.com/abstract=2785970.

Matthew Gentzkow and Emir Kamenica. A Rothschild-Stiglitz Approach to Bayesian Persuasion. *American Economic Review*, 106(5):597–601, May 2016. ISSN 0002-8282. doi: 10.1257/aer.p20161049. URL https://www.aeaweb.org/articles?id=10.1257/aer.p20161049.

Anton Kolotilin. Optimal Information Disclosure: A Linear Programming Approach. SSRN Scholarly Paper ID 2866121, Social Science Research Network, Rochester, NY, October 2016. URL https://papers.ssrn.com/abstract=2866121.

Volker Strassen. The Existence of Probability Measures with Given Marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965. ISSN 0003-4851. URL http://www.jstor.org/stable/2238148.