This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Electronic companion for "Column-Randomized Linear Programs: Performance Guarantees and Applications" by Akchen and Mišić

Table of Contents

A Additional Discussion on Results in Section 3 **ec2**

B Omitted Proofs of Results in Section 3 **ec8**

C Omitted Proofs and Other Results for Section 4 **ec16**

D Special Structures and Extensions **ec35**

E Statistically-Dependent Columns **ec47**

F Cutting Stock Problem Experiments (continued) **ec57**

G Nonparametric Choice Model Estimation Experiments (continued) **ec64**

H Comparisons to Other Approaches **ec68**

The ecompanion is organized as follows. Section A follows Section 3.3 and continues the discussion of Theorems 1 and 2. Section B completes all omitted proofs for the theoretical results in Section 3. Section C provides the proofs for all theoretical results in Section 4 regarding the distributional counterpart and provides additional comments. Section D shows that the parameters γ in Theorem 1 can be further obtained for several applications. Section E extends the proposed framework by proposing sampling statistically-dependent columns in the column randomization method. Section F follows Section 5 and completes the numerical experiments on the cutting stock problem. Section G follows Section 6 and completes the numerical experiments on the nonparametric choice model estimation. Section H provides a detailed comparison between the proposed framework and other large-scale LP solvers based on randomized algorithms (Agrawal et al. 2014, Vu et al. 2018). All of the code is available at the repository `Column-Randomized_LP` at

`https://github.com/yi-chun-akchen/Column-Randomized_LP`.

Appendix A Additional Discussion on Results in Section 3

This section continues the discussion in Section 3.3.

Lower Bound on $v(P_J)$: We note that neither Theorem 1 nor 2 implies that the optimality gap $\Delta v(P_J)$ of the column-randomized linear program P_J can be arbitrarily small with large K. Indeed, if $\boldsymbol{\xi}$ is not "comprehensive" enough – that is, its support is small, and does not include the complete set of columns of any optimal basis for P – then $\Delta v(P_{\text{distr}}) > 0$ no matter what C is, and one would not expect the column-randomized program P_J to perform closely to the complete problem P, even if K is large. We can formalize this intuition in the following proposition, where I^+ denotes the support of the distribution $\boldsymbol{\xi}$.

PROPOSITION EC.1. *Define* $I^+ = \{j \in [n] \mid \xi_j > 0\}$ *and let* $P^+_{distr} \equiv \min\{\mathbf{c}^T_{I^+}\mathbf{x}^+ \mid \mathbf{A}_{I^+}\mathbf{x}^+ = \mathbf{b}, \mathbf{x}^+ \geq \mathbf{0}\}$. *Then* $v(P_J) \geq v(P^+_{distr})$ *almost surely, and* $v(P_J) \to v(P^+_{distr})$ *almost surely as* $K \to \infty$.

The proof is straightforward and omitted for brevity, as any solution of P_J can be reformulated as a feasible solution of P^+_{distr}, and as $K \to \infty$, every column in I^+ is sampled at least once almost surely, ensuring that $v(P_J) = v(P^+_{\text{distr}})$. An obvious consequence of this proposition is that if $I^+ = [n]$, i.e., every column has a positive probability of being sampled, then $v(P_J)$ will converge to $v(P)$ when enough columns are sampled. From this perspective, the value of our bounds in Theorems 1 and 2 is that they provide finite sample guarantees, for the case where $K \ll n$ and it is impossible that one will have sampled all of the n columns.

Feasibility of P_J: We make several important remarks regarding the feasibility of P_J and how feasibility is incorporated in our guarantee. First, note that in general, the sampled problem P_J need not be feasible. As a simple example, consider the following complete problem:

$$P = P_{\mathbf{I}} \equiv \min\{\mathbf{1}^T\mathbf{x} \mid \mathbf{I}\mathbf{x} = \mathbf{1}, \mathbf{x} \geq \mathbf{0}\},$$

where $\mathbf{I}$ is the n-by-n identity matrix and $m = n$. In this problem, the only way that the sampled problem P_J can be feasible is if the collection $j_1, \ldots, j_K$ includes every index in $[n]$; if any column $j \in [n]$ is not part of the sample J, then the sampled problem P_J is automatically infeasible. Thus, when $K < n$, P_J is infeasible almost surely. When $K \geq n$, it is still possible that $j_1, \ldots, j_K$ does not include all indices in $[n]$, and thus P_J is infeasible with positive probability.

For this reason, our guarantee on the optimality gap is stated as a *conditional* guarantee: with high probability over the sample $j_1, \ldots, j_K$, the optimality gap of P_J obeys a particular bound *if the column-randomized LP is feasible*. Formally, our two guarantees can be represented as

$$\mathbf{Pr}\left[\{P_J \text{ is feasible}\} \Rightarrow \left\{\Delta v(P_J) \leq \Delta v(P_{\text{distr}}) + \frac{C}{\sqrt{K}} \cdot C_P \cdot C_\delta\right\}\right] \geq 1 - \delta.$$

Since the implication $A \Rightarrow B$ is logically equivalent to $A^C \cup B$, an alternative equivalent restatement of the general form of our guarantee is

$$\mathbf{Pr}\left[\{P_J \text{ is infeasible}\} \cup \left\{\Delta v(P_J) \leq \Delta v(P_{\text{distr}}) + \frac{C}{\sqrt{K}} \cdot C_P \cdot C_\delta\right\}\right] \geq 1 - \delta.$$

We note that this type of guarantee is distinct from probabilistically conditioning on $j_1, \ldots, j_K$, i.e., our guarantee is *not* the same as

$$\mathbf{Pr}\left[\Delta v(P_J) \leq \Delta v(P_{\text{distr}}) + \frac{C}{\sqrt{K}} \cdot C_P \cdot C_\delta \;\middle|\; P_J \text{ is feasible}\right] \geq 1 - \delta,$$

because upon conditioning on the feasibility of P_J, the random variables $j_1, \ldots, j_K$ are in general no longer an i.i.d. sample. As an example of this, consider again problem $P_{\mathbf{I}}$ above, with $K = n$ and a randomization scheme ρ corresponding to the uniform distribution $\boldsymbol{\xi} = (1/n, \ldots, 1/n)$ over $[n]$. By conditioning on the event that P_J is feasible, the sample $J = \{j_1, \ldots, j_K\}$ must then be exactly equal to $[n]$, and we obtain that $\mathbf{Pr}[j_k = t, j_{k'} = t] = 0 \neq \mathbf{Pr}[j_k = t] \cdot \mathbf{Pr}[j_{k'} = t]$ for any $k, k' \in [K]$ with $k \neq k'$ and $t \in [n]$. In this example, the indices $j_1, \ldots, j_K$ are thus not independent.

With regard to the feasibility of column-randomized LPs, it appears to be difficult to guarantee feasibility in general. However, one can use similar techniques as in the proofs of our main results to characterize the near-feasibility of a column-randomized LP. Consider the following complete problem, and its sampled and distributional counterparts:

$$\begin{aligned}
P^{\text{feas}} &= \min\{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 \mid \mathbf{x} \geq \mathbf{0}\}, \\
P_J^{\text{feas}} &= \min\{\|\mathbf{A}_J\tilde{\mathbf{x}} - \mathbf{b}\|_1 \mid \tilde{\mathbf{x}} \geq \mathbf{0}\}, \\
P_{\text{distr}}^{\text{feas}} &= \min\{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1 \mid \mathbf{0} \leq \mathbf{x} \leq C\boldsymbol{\xi}\}.
\end{aligned}$$

The objective function in each problem measures how close $\mathbf{A}\mathbf{x}$ is to $\mathbf{b}$ for a given nonnegative solution $\mathbf{x}$, and the optimal value measures the minimum total infeasibility, as measured by the lowest attainable ℓ_1 distance between $\mathbf{A}\mathbf{x}$ and $\mathbf{b}$. Note that an optimal value of zero for a given problem implies that the feasible region contains a solution $\mathbf{x}$ that satisfies $\mathbf{A}\mathbf{x} = \mathbf{b}$. With a slight abuse of notation, let us use $v(P^{\text{feas}})$, $v(P_J^{\text{feas}})$ and $v(P_{\text{distr}}^{\text{feas}})$ to denote the optimal objective value of each problem. We then have the following result.

PROPOSITION EC.2. *Let C be a nonnegative constant. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the sample J,*

$$v(P_J^{feas}) \leq v(P_{distr}^{feas}) + \frac{C}{\sqrt{K}} \cdot m \cdot \|\mathbf{A}\|_{\max} \cdot \left(1 + \sqrt{2 \log \frac{1}{\delta}}\right).$$

The proof of Proposition EC.2 (see Section B.3 of the ecompanion) follows using a similar but simpler procedure than those used in the proofs of Theorems 1 and 2. The guarantee in Proposition EC.2 has a similar interpretation to Theorems 1 and 2: the magnitude of the total infeasibility of the columns J is bounded with high probability by the minimum infeasibility of the distributional counterpart $P_{\text{distr}}^{\text{feas}}$ plus a $O(1/\sqrt{K})$ term.

Feasibility-guaranteed column randomization algorithm: One practical way in which one can modify Algorithm 1 to ensure that the sampled problem is always feasible is to augment the column set J with a set of columns J_F such that $P_{J\cup J_F}$ is a feasible problem. We define the new procedure, Algorithm 7, below.

Algorithm 7 The Feasibility Guaranteed Column Randomization Method

1: Set $J_F \subseteq [n]$ to be a set of columns such that P_{J_F} is feasible and $\text{rank}(\mathbf{A}_{J_F}) = m$.
2: Sample K indices as $J \equiv \{j_1, \ldots, j_K\}$ by a randomization scheme ρ.
3: Define $\mathbf{A}_{J\cup J_F} = [A_j]_{j\in J\cup J_F}$ and $\mathbf{c}_{J\cup J_F} = [c_j]_{j\in J\cup J_F}$.
4: Solve the column-randomized linear program, which has $|J \cup J_F|$ columns:

$$P_{J\cup J_F}: \quad \min\left\{\mathbf{c}_{J\cup J_F}^T \tilde{\mathbf{x}} \,|\, \mathbf{A}_{J\cup J_F}\tilde{\mathbf{x}} = \mathbf{b},\ \tilde{\mathbf{x}} \geq \mathbf{0}\right\}. \tag{EC.1}$$

5: **return** optimal objective value $v(P_{J\cup J_F})$ and an optimal solution $\tilde{\mathbf{x}}^*$.

There are two important aspects of Algorithm 7 to be cognizant of. First, for this procedure, we can adapt Theorems EC.1 and Theorem EC.2 so as to obtain guarantees on $\Delta v(P_{J\cup J_F})$. This results in the following two guarantees; importantly, these guarantees are no longer conditional guarantees.

THEOREM EC.1. *Let J and J_F be as defined in Algorithm 7. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, then*

$$\Delta v(P_{J\cup J_F}) \leq \Delta v(P_{distr}) + \frac{C\,(1 + m\gamma\|\mathbf{A}\|_{\max})}{\sqrt{K}}\left(1 + \sqrt{2\log\frac{2}{\delta}}\right), \tag{EC.2}$$

THEOREM EC.2. *Let J and J_F be as defined in Algorithm 7. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, then*

$$\Delta v(P_{J\cup J_F}) \leq \Delta v(P_{distr}) + \frac{C}{\sqrt{K}} \cdot \chi \cdot \left(1 + \sqrt{2\log\frac{1}{\delta}}\right) \tag{EC.3}$$

Second, in the statement of Algorithm 7, the first step is to obtain a set of columns J_F so that P_{J_F} is feasible and $\mathbf{A}_{J_F}$ has full row rank. This can be accomplished easily by applying the first phase of the two-phase method, which is a standard method for obtaining an initial feasible solution to a linear program for which an initial basis is not obvious (see Chapter 3, Section 5 of Bertsimas and Tsitsiklis 1997). In particular, one formulates the following problem:

$$\begin{aligned}
\underset{\mathbf{x},\boldsymbol{\epsilon}^+,\boldsymbol{\epsilon}^-}{\text{minimize}} \quad & \mathbf{1}^T\boldsymbol{\epsilon}^+ + \mathbf{1}^T\boldsymbol{\epsilon}^- && \text{(EC.4a)}\\
\text{subject to} \quad & \mathbf{A}\mathbf{x} + \mathbf{I}\boldsymbol{\epsilon}^+ - \mathbf{I}\boldsymbol{\epsilon}^- = \mathbf{b}, && \text{(EC.4b)}\\
& \mathbf{x} \geq \mathbf{0}, && \text{(EC.4c)}\\
& \boldsymbol{\epsilon}^+, \boldsymbol{\epsilon}^- && \text{(EC.4d)}
\end{aligned}$$

where $\boldsymbol{\epsilon}^+, \boldsymbol{\epsilon}^- \in \mathbb{R}^m$ and $\mathbf{0}$ is an appropriately sized vector of zeros in the two nonnegativity constraints. As with the main problem P, this problem obviously cannot be formulated explicitly, but can be solved using column generation. At the start of column generation, we do not include any columns from $\mathbf{A}$, and the principal constraint is $\mathbf{I}\boldsymbol{\epsilon}^+ - \mathbf{I}\boldsymbol{\epsilon}^- = \mathbf{b}$, for which an initial basis can be found trivially: for each $i \in [m]$, set $\epsilon_i^+ = \max\{b_i, 0\}$ and $\epsilon_i^- = \min\{b_i, 0\}$. As column generation progresses, the objective value will decrease and the ϵ_i^+ and ϵ_i^- variables will gradually leave the basis. Upon termination, one will obtain a basic feasible solution for which all of the $\epsilon_i^+, \epsilon_i^-$ variables are non-basic. The resulting set of basic columns of $\mathbf{x}$, J_F, is such that P_{J_F} is feasible and $\text{rank}(\mathbf{A}_{J_F}) = m$. Observe now that if P_{J_F} is feasible, then $P_{J \cup J_F}$, for *any* set of columns $J \subseteq [n]$, must also be feasible: one can take any solution $\mathbf{x}_{J_F}$ to P_{J_F} and set $x_j = 0$ for all $j \in J \setminus J_F$, resulting in a solution $\mathbf{x}_{J \cup J_F}$ that is nonnegative, and satisfies $\mathbf{A}_{J \cup J_F}\mathbf{x}_{J \cup J_F} = \mathbf{b}$. In addition, the matrix $\mathbf{A}_{J \cup J_F}$ must also have rank m. Although this approach requires column generation, it is reasonable to expect that column generation applied to problem (EC.4) to find an initial solution should generally be faster than when it is applied to the complete problem P. Proofs of Theorem EC.1 and EC.2 can be found in Section B.4.

Interpretation of γ and χ: We first note that the technique of bounding the objective value of a linear program using the ℓ_∞ norm of basic feasible solutions has been applied previously in the literature (Ye 2011, Kitahara and Mizuno 2013). The presence of γ and χ in Theorem 1 and 2, respectively, arises due to the use of sensitivity analysis results from linear programming with respect to the right-hand side vector $\mathbf{b}$. As we discuss in the proof in Section B, any optimal solution $\mathbf{x}^{*0}$ of problem P_{distr} has a sparse counterpart $\mathbf{x}'$ in the space $\mathcal{S}_J \equiv \{\mathbf{x} \mid x_j = 0 \ \forall j \notin J\}$ such that $\mathbf{x}'$ is in the vicinity of $\mathbf{x}^{*0}$ in terms of Euclidean distance. However, $\mathbf{x}'$ does not necessarily belong to the feasible set $\mathcal{F}(P_J)$ of the column-randomized linear program P_J, since $\mathcal{F}(P_J)$ is a subset of $\mathcal{S}_J$. To relate the optimal objective value $v(P_J)$ of problem P_J to $\mathbf{c}^T\mathbf{x}'$, which is close to $\mathbf{c}^T\mathbf{x}^{*0}$, we use sensitivity analysis arguments which involve either γ or χ.

Comparison of Theorems 1 and 2: While both Theorem 1 and 2 provide valid bounds for the optimality gap $\Delta v(P_J)$, Theorem 1 is in general easier to apply; indeed, in Section D we discuss two notable examples where γ can be easily computed (specifically, LPs with totally unimodular constraint matrices $\mathbf{A}$ and infinite horizon discounted Markov decision processes). For problems that are not standard form LPs, neither guarantee directly applies, but we can obtain specialized guarantees by carefully modifying a result (Proposition EC.3 in Section B.2) that leads to Theorem 1 and designing bounds for the ℓ_∞ norm of feasible or optimal solutions of D_J (as opposed to basic solutions of D). We will later showcase two examples of such guarantees, for covering LPs (Section D.3) and packing LPs (Section D.4).

With regard to Theorem 2, we expect for most problems that Theorem 2 will be difficult to apply, as it requires a universal bound for the norm of the reduced cost vector for every basis, feasible or not, of problem P. Nevertheless, Theorem 2 is interesting because it involves reduced costs, which are also of importance in column generation. For a basic feasible solution, the reduced cost of a non-basic variable j can be thought of as the rate at which the objective changes as one increases x_j to move from the current basic feasible solution to an adjacent/neighboring basic feasible solution in which j is part of the basis. With this perspective of reduced costs, one can informally interpret the result in the following way: if χ is small, then the rate at which the objective changes between adjacent basic feasible solutions is small. In such a setting, it is reasonable to expect that there will be many basic feasible solutions that are close to being optimal and that solving the sampled problem P_J should return a solution that performs well. On the other hand, if there exist non-optimal basic feasible solutions where the reduced cost vector has a very large magnitude (which would imply a large χ), then this would suggest that the objective changes by a large amount between certain adjacent basic feasible solutions, and that there are certain "good" columns that are more important than others for achieving a low objective value. In this setting, we would expect the sampled problem objective $v(P_J)$ to only be close to $v(P)$ if J includes the "good" columns, which would be unlikely to happen in general.

Design of Randomization Scheme ρ: The quantity ξ_j, which is the probability that the jth column is drawn by the randomization scheme ρ, can be interpreted as the relative importance of x_j compared to other components of $\mathbf{x} \in \mathbb{R}^n$ in the complete problem P; indeed, when the corresponding column is randomly chosen, x_j is allowed to be nonzero, and can thus be utilized to solve the optimization problem. For example, in a network flow optimization problem, x_j represents the amount of flow over edge j; a nonzero ξ_j can thus be interpreted as the belief that edge j should be used for flow. As another example, consider the LP formulation of an MDP, where each component of $\mathbf{x}$ corresponds to a state-action pair (s, a) (i.e., $x_{(s,a)}$ is the expected discounted frequency of the system being in state s and action a being taken). In this setting, a nonzero $\xi_{(s,a)}$ can be interpreted as the relative importance of (s, a) to other state-action pairs.

One can design the randomization scheme based on prior knowledge of the problem. For example, one could use a heuristic solution to a network flow problem to design a randomization scheme ρ resulting in a distribution $\boldsymbol{\xi}$ that is biased towards this heuristic solution. Similarly, if one has access to a good heuristic policy for an MDP, one can design a distribution $\boldsymbol{\xi}$ that is biased towards state-action pairs (s, a) that occur frequently for this policy. If such prior knowledge is not available, a uniform or nearly-uniform distribution over $[n]$ is adequate. We provide several concrete examples on how to design randomization schemes in our numerical experiments in Sections 5 and 6. Finally, we note that the indices in J have been assumed to be i.i.d. In Section E, we derive analogous guarantees for the case when the indices are sampled non-independently.

Minor Remarks on the Upper Bound: We mention two other interesting properties of the bound (8). First, the second term in (8) is independent of the distribution $\boldsymbol{\xi}$; no matter how $\boldsymbol{\xi}$ is designed, the optimality gap $\Delta v(P_J)$ is guaranteed to converge with rate $1/\sqrt{K}$. Second, the dependence of the bound on the confidence parameter δ is via $\sqrt{2\log(2/\delta)}$ in Theorem 1 or $\sqrt{2\log(1/\delta)}$ in Theorem 2. This implies that very small values of δ will not significantly increase the upper bound on $\Delta v(P_J)$.

Appendix B Omitted Proofs of Results in Section 3

In this section, we prove Theorem 1 and 2. We start with some preliminary results (Section B.1) then prove the main theorems (Section B.2).

B.1 Preliminary Results and Lemmas

Lemma EC.1 and EC.2 bound the distance between the sample mean and the expected value of a collection of i.i.d. vectors, in terms of ℓ_2 norm and ℓ_1 norm, respectively. Lemma EC.1 is Lemma 4 from Rahimi and Recht (2009), which utilizes McDiarmid's inequality to show that the scalar function $\|\bar{\mathbf{w}} - \mathbb{E}[\bar{\mathbf{w}}]\|_2$, where $\bar{\mathbf{w}}$ is the mean of K i.i.d. vectors $\mathbf{w}_1, \ldots, \mathbf{w}_K$, concentrates to zero with rate $O\left(1/\sqrt{K}\right)$.

LEMMA EC.1. *(Rahimi and Recht 2009) Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$ be i.i.d. random vectors such that $\|\mathbf{w}_k\|_2 \le C$ for $k = 1, \ldots, K$. Let $\bar{\mathbf{w}} = (1/K) \cdot \sum_{k=1}^{K} \mathbf{w}_k$. Then for any $\delta \in (0,1)$, we have, with probability at least $1-\delta$,*

$$\|\bar{\mathbf{w}} - \mathbb{E}[\bar{\mathbf{w}}]\|_2 \le \frac{C}{\sqrt{K}} \cdot \left(1 + \sqrt{2\log\frac{1}{\delta}}\right).$$

LEMMA EC.2. *Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$ be i.i.d. random vectors of size m such that $\|\mathbf{w}_k\|_\infty \le C$ for $k = 1, \ldots, K$. Let $\bar{\mathbf{w}} = (1/K) \cdot \sum_{k=1}^{K} \mathbf{w}_k$. Then for any $\delta \in (0,1)$, we have, with probability at least $1-\delta$,*

$$\|\bar{\mathbf{w}} - \mathbb{E}[\bar{\mathbf{w}}]\|_1 \le \frac{mC}{\sqrt{K}} \cdot \left(1 + \sqrt{2\log\frac{1}{\delta}}\right).$$

Proof: Since $\|\mathbf{w}_k\|_2 \le \sqrt{m}\|\mathbf{w}_k\|_\infty \le \sqrt{m}C$, we apply Lemma EC.1 and obtain that with probability at least $1-\delta$, $\|\bar{\mathbf{w}} - \mathbb{E}[\bar{\mathbf{w}}]\|_2 \le \sqrt{m} \cdot C/\sqrt{K} \cdot \left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$. Combining this with the fact that $\|\bar{\mathbf{w}} - \mathbb{E}[\bar{\mathbf{w}}]\|_1 \le \sqrt{m} \cdot \|\bar{\mathbf{w}} - \mathbb{E}[\bar{\mathbf{w}}]\|_2$, we obtain the desired result. □

Lemma EC.3 is a standard result of sensitivity analysis of linear programming; see Chapter 5 of Bertsimas and Tsitsiklis (1997). In fact, one can view the optimal objective value of problem P as a convex function in $\mathbf{b}$ and show that any optimal dual solution $\mathbf{p}$ is a subgradient at $\mathbf{b}$.

LEMMA EC.3. *Let $z(\mathbf{b}) = \min\{\mathbf{c}_0^T\mathbf{y} \mid \mathbf{A}_0\mathbf{y} = \mathbf{b}, \mathbf{y} \ge \mathbf{0}\}$ and $z(\mathbf{b}') = \min\{\mathbf{c}_0^T\mathbf{y} \mid \mathbf{A}_0\mathbf{y} = \mathbf{b}', \mathbf{y} \ge \mathbf{0}\}$. Then $z(\mathbf{b}) - z(\mathbf{b}') \le \mathbf{p}^T(\mathbf{b} - \mathbf{b}')$, where $\mathbf{p}$ is an optimal dual solution of the former problem.*

B.2 Proofs of Theorem 1 and 2

We first establish a useful result.

PROPOSITION EC.3. *Let C be a nonnegative constant and define the linear program P_{distr} as in Theorem 1, i.e., $P_{distr}:\ \min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x}=\mathbf{b}, \mathbf{0}\leq\mathbf{x}\leq C\boldsymbol{\xi}\}$. Let P_J be the column-randomized LP solved by Algorithm 1. For any $\delta\in(0,1)$, with probability at least $1-\delta$ over the sample J, the following holds: if P_J is feasible, then*

$$\Delta v(P_J)\leq \Delta v(P_{distr}) + \frac{C}{\sqrt{K}}\cdot(1+\|\mathbf{p}\|_\infty\cdot m\cdot\|\mathbf{A}\|_{\max})\cdot\left(1+\sqrt{2\log\frac{2}{\delta}}\right)$$

for any optimal solution $\mathbf{p}$ of problem D_J (the dual of problem P_J).

Proof: Let $j_1,\ldots,j_K$ be the set of indices sampled according to the distribution $\boldsymbol{\xi}$ by the randomization scheme ρ. Let $\mathbf{x}^{*0}$ be an optimal solution of the distributional counterpart problem P_{distr}. Consider the solution $\mathbf{x}'$ that is defined as

$$\mathbf{x}' \equiv \frac{1}{K}\sum_{k=1}^{K}\frac{x^{*0}_{j_k}}{\xi_{j_k}}\cdot\mathbf{e}_{j_k},$$

where we use $\mathbf{e}_j$ to denote the jth standard basis vector for $\mathbb{R}^n$. In addition, define the vector $\mathbf{b}'$ as

$$\mathbf{b}' \equiv \mathbf{A}\mathbf{x}'.$$

To prove our result, we proceed in three steps. In the first step, we show how we can probabilistically bound $\|\mathbf{x}'-\mathbf{x}^{*0}\|_2$. In the second step, we show how we can probabilistically bound $\|\mathbf{b}'-\mathbf{b}\|_1$. In the last step, we use the results of our first two steps, together with sensitivity results for linear programs, to derive the required bound. In what follows, we use I_+ to denote the support of $\boldsymbol{\xi}$, that is, $I_+=\{j\in[n]\mid\xi_j>0\}$.

Step 1: Bounding $\|\mathbf{x}'-\mathbf{x}^{*0}\|_2$. To show that $\mathbf{x}'$ will be close to $\mathbf{x}^{*0}$, let us first define the vector $\mathbf{w}_k$ as

$$\mathbf{w}_k = \frac{x^{*0}_{j_k}}{\xi_{j_k}}\cdot\mathbf{e}_{j_k}$$

for each $k\in[K]$. The vectors $\mathbf{w}_1,\ldots,\mathbf{w}_K$ constitute an i.i.d. collection of vectors, and possess three special properties. First, observe that $\mathbf{x}'$ is just the sample mean of $\mathbf{w}_1,\ldots,\mathbf{w}_K$. Second, observe that the expected value of each $\mathbf{w}_k$ can be calculated as

$$\mathbb{E}[\mathbf{w}] = \sum_{j\in I_+}\xi_j\cdot\frac{x^{*0}_j}{\xi_j}\cdot\mathbf{e}_j = \sum_{j\in I_+}x^{*0}_j\mathbf{e}_j = \sum_{j\in[n]}x^{*0}_j\mathbf{e}_j = \mathbf{x}^{*0}$$

where we use $\mathbf{w}$ to denote a random vector following the same distribution as each $\mathbf{w}_k$. In the above, we note that the third step follows because the distributional counterpart P_{distr} includes the constraint $\mathbf{x} \leq C\boldsymbol{\xi}$, so $j \notin I_+$ automatically implies that $x_j^{*0} = 0$.

Finally, observe that the ℓ_2 norm of each $\mathbf{w}_k$ can be bounded as

$$\|\mathbf{w}_k\|_2 = \left|\frac{x_{j_k}^{*0}}{\xi_{j_k}}\right| \cdot \|\mathbf{e}_{j_k}\|_2 \leq C \cdot 1 = C,$$

where the inequality follows because $\mathbf{x}^{*0}$ satisfies the constraint $\mathbf{0} \leq \mathbf{x} \leq C\boldsymbol{\xi}$. With these three properties in hand, and recognizing that $\|\mathbf{x}' - \mathbf{x}^{*0}\|_2 = \|(1/K)\sum_{k=1}^K \mathbf{w}_k - \mathbb{E}[\mathbf{w}]\|_2$, we can invoke Lemma EC.1 to assert that, with probability at least $1 - \delta/2$,

$$\|\mathbf{x}' - \mathbf{x}^{*0}\|_2 \leq \frac{C}{\sqrt{K}} \cdot \left(1 + \sqrt{2\log\frac{2}{\delta}}\right). \tag{EC.5}$$

Step 2: Bounding $\|\mathbf{b}' - \mathbf{b}\|_1$. To show that $\mathbf{b}'$ will be close $\mathbf{b}$, we proceed similarly to Step 1. In particular, we define $\mathbf{b}_k$ for each $k \in [K]$ as

$$\mathbf{b}_k \equiv \mathbf{A}\mathbf{w}_k = \frac{x_{j_k}^{*0}}{\xi_{j_k}} \cdot \mathbf{A}\mathbf{e}_{j_k} = \frac{x_{j_k}^{*0}}{\xi_{j_k}}\mathbf{A}_{j_k}.$$

Observe that by definition of $\mathbf{b}_k$, we have that the sample mean of $\mathbf{b}_1, \ldots, \mathbf{b}_K$ is equal to $\mathbf{b}'$:

$$\frac{1}{K}\sum_{k=1}^K \mathbf{b}_k = \frac{1}{K}\sum_{k=1}^K \mathbf{A}\mathbf{w}_k = \mathbf{A}\left(\frac{1}{K}\sum_{k=1}^K \mathbf{w}_k\right) = \mathbf{A}\mathbf{x}' \equiv \mathbf{b}'. \tag{EC.6}$$

In addition, the expected value of each $\mathbf{b}_k$ can be calculated; letting $\tilde{\mathbf{b}}$ denote a random variable with the same distribution as each $\mathbf{b}_k$, we have

$$\mathbb{E}[\tilde{\mathbf{b}}] = \mathbf{A}\mathbb{E}[\mathbf{w}_k] = \mathbf{A}\mathbf{x}^{*0} = \mathbf{b}.$$

Lastly, we can bound the ℓ_∞ norm of each vector $\mathbf{b}_k$ as

$$\|\mathbf{b}_k\|_\infty = \left\|\frac{x_{j_k}^{*0}}{\xi_{j_k}}\mathbf{A}_{j_k}\right\|_\infty = \left|\frac{x_{j_k}^{*0}}{\xi_{j_k}}\right| \cdot \|\mathbf{A}_{j_k}\|_\infty \leq C\|\mathbf{A}\|_{\max},$$

where the inequality follows by the definition of $\|\mathbf{A}\|_{\max}$ and the fact that $\mathbf{x}^{*0}$ satisfies $\mathbf{0} \leq \mathbf{x} \leq C\boldsymbol{\xi}$.

With these observations in hand, we now recognize that $\|\mathbf{b}' - \mathbf{b}\|_1 = \|(1/K)\sum_{k=1}^K \mathbf{b}_k - \mathbb{E}[\tilde{\mathbf{b}}]\|_1$, i.e., $\|\mathbf{b}' - \mathbf{b}\|_1$ is just the ℓ_1 norm of the deviation of a sample mean from its true expectation; we can therefore invoke Lemma EC.2 to assert that, with probability at least $1 - \delta/2$,

$$\|\mathbf{b}' - \mathbf{b}\|_1 \leq \frac{m \cdot C \cdot \|\mathbf{A}\|_{\max}}{\sqrt{K}} \cdot \left(1 + \sqrt{2\log\frac{2}{\delta}}\right). \tag{EC.7}$$

Step 3: Completing the proof. With Steps 1 and 2 complete, we are now ready to bound the optimality gap. For any vector $\mathbf{b}'' \in \mathbb{R}^m$, we define the linear program $P_J(\mathbf{b}'')$ as

$$P_J(\mathbf{b}''): \quad \min\left\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}'', \mathbf{x} \geq \mathbf{0},\ x_j = 0\ \forall j \notin J\right\}. \tag{EC.8}$$

Then $v(P_J(\mathbf{b}')) \leq \mathbf{c}^T\mathbf{x}'$; this follows because $\mathbf{A}\mathbf{x}' = \mathbf{b}'$ and $\mathbf{x}' \geq \mathbf{0}$, which means that $\mathbf{x}'$ is a feasible solution to problem $P_J(\mathbf{b}')$. In addition, since $\mathbf{c}^T\mathbf{x}^{*0} = v(P_{\text{distr}})$, we have

$$v(P_J(\mathbf{b}')) \leq \mathbf{c}^T\mathbf{x}' = \mathbf{c}^T\left(\mathbf{x}^{*0} + (\mathbf{x}' - \mathbf{x}^{*0})\right) = v(P_{\text{distr}}) + \mathbf{c}^T(\mathbf{x}' - \mathbf{x}^{*0}). \tag{EC.9}$$

If the column-randomized problem P_J is feasible, then by letting $\mathbf{p}$ be any optimal solution of the dual of P_J and applying Lemma EC.3, we have

$$\begin{aligned}
v(P_J) = v(P_J(\mathbf{b})) &\leq v(P_J(\mathbf{b}')) + \mathbf{p}^T(\mathbf{b} - \mathbf{b}') && \text{(EC.10)}\\
&\leq v(P_{\text{distr}}) + \mathbf{c}^T(\mathbf{x}' - \mathbf{x}^{*0}) + \mathbf{p}^T(\mathbf{b} - \mathbf{b}') && \text{(EC.11)}\\
&\leq v(P_{\text{distr}}) + \|\mathbf{c}\|_2 \cdot \|\mathbf{x}' - \mathbf{x}^{*0}\| + \|\mathbf{p}\|_\infty \cdot \|\mathbf{b}' - \mathbf{b}\|_1 && \text{(EC.12)}\\
&= v(P_{\text{distr}}) + \|\mathbf{x}' - \mathbf{x}^{*0}\|_2 + \|\mathbf{p}\|_\infty \cdot \|\mathbf{b}' - \mathbf{b}\|_1, && \text{(EC.13)}
\end{aligned}$$

where the first inequality comes from Lemma EC.3, the second inequality comes from (EC.9), the third inequality comes from the Cauchy-Schwarz inequality and Hölder's inequality, and the last equality comes from the assumption that $\|\mathbf{c}\|_2 = 1$.

We now bound expression (EC.13) by applying the inequalities (EC.5) and (EC.7), each of which hold with probability at least $1 - \delta/2$, and combining them using the union bound. We thus obtain that, with probability at least $1 - \delta$,

$$v(P_J) \leq v(P_{\text{distr}}) + \frac{C}{\sqrt{K}} \cdot (1 + \|\mathbf{p}\|_\infty \cdot m \cdot \mathbf{A}_{\max}) \cdot \left(1 + \sqrt{2\log\frac{2}{\delta}}\right). \tag{EC.14}$$

Subtracting $v(P)$ from both sides gives us the required inequality. □

With Proposition EC.3, we can smoothly prove Theorem 1 as follows.

Proof of Theorem 1: By invoking Proposition EC.3, we obtain that with probability at least $1 - \delta$, if P_J is feasible, then

$$\Delta v(P_J) \leq \Delta v(P_{\text{distr}}) + \frac{C}{\sqrt{K}} \cdot (1 + \|\mathbf{p}\|_\infty \cdot m \cdot \mathbf{A}_{\max}) \cdot \left(1 + \sqrt{2\log\frac{2}{\delta}}\right),$$

for any dual optimal solution $\mathbf{p}$ of D_J. To prove the theorem, let us set $\mathbf{p}$ to an optimal basic feasible solution of the problem D_J. Note that such a dual optimal solution is guaranteed to exist by the assumption that $\text{rank}(\mathbf{A}_J) = m$. Since $\mathbf{p}$ is a basic feasible solution of D_J, it is automatically a basic (but not necessarily feasible) solution of the complete dual problem D. By the definition of γ in the theorem, we have that $\|\mathbf{p}\|_\infty \leq \gamma$, and the theorem follows. □

To prove Theorem 2, we prove a complementary result to Proposition EC.3.

PROPOSITION EC.4. *Let C, P_J and P_{distr} be defined as in the statement of Proposition EC.3. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, the following holds: if P_J is feasible, then*

$$\Delta v(P_J) \leq \Delta v(P_{distr}) + \frac{C}{\sqrt{K}} \cdot \|\mathbf{c}^T - \mathbf{p}^T\mathbf{A}\|_2 \cdot \left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$$

for any optimal solution $\mathbf{p}$ of problem D_J (the dual of problem P_J).

Proof: We follow the proof of Proposition EC.3 until inequality (EC.11) and continue as follows:

$$\begin{aligned} v(P_J) = v(P_J(\mathbf{b})) &\leq v(P_J(\mathbf{b}')) + \mathbf{p}^T(\mathbf{b}-\mathbf{b}') \\ &\leq v(P_{\text{distr}}) + \mathbf{c}^T(\mathbf{x}'-\mathbf{x}^{*0}) + \mathbf{p}^T(\mathbf{b}-\mathbf{b}') \\ &= v(P_{\text{distr}}) + \mathbf{c}^T(\mathbf{x}'-\mathbf{x}^{*0}) + \mathbf{p}^T\mathbf{A}(\mathbf{x}^{*0}-\mathbf{x}') \\ &= v(P_{\text{distr}}) + \left(\mathbf{c}^T - \mathbf{p}^T\mathbf{A}\right)(\mathbf{x}'-\mathbf{x}^{*0}) \\ &\leq v(P_{\text{distr}}) + \|\mathbf{c}^T - \mathbf{p}^T\mathbf{A}\|_2 \cdot \|\mathbf{x}'-\mathbf{x}^{*0}\|_2, \end{aligned} \tag{EC.15}$$

where the bound holds for any optimal solution $\mathbf{p}$ of the sampled dual problem D_J. By invoking Lemma EC.1 with δ to bound $\|\mathbf{x}'-\mathbf{x}^{*0}\|_2$, and subtracting $v(P)$ from both sides, we obtain the desired result. □

Using Proposition EC.4, we now prove Theorem 2.

Proof of Theorem 2: We invoke Proposition EC.4 and set $\mathbf{p}$ to be an optimal basic feasible solution of the sampled dual problem D_J; then $\mathbf{p}^T = \mathbf{c}_B^T\mathbf{A}_B^{-1}$ for some set of basic variables $B \subset [n]$. In this case, we observe that the dual slack vector $\mathbf{c}^T - \mathbf{p}^T\mathbf{A}$ becomes $\mathbf{c}^T - \mathbf{c}_B^T\mathbf{A}_B^{-1}\mathbf{A}$, which is exactly the reduced cost vector $\bar{\mathbf{c}}$ associated with the basis B within the full problem P. By using the hypothesis that any such reduced cost vector satisfies $\|\bar{\mathbf{c}}\|_2 \leq \chi$, we obtain the desired result. □

B.3 Proof of Proposition EC.2

Let $\mathbf{x}^{*0}$ be an optimal solution of $P^{\text{feas}}_{\text{distr}}$. Define the solution $\mathbf{x}'$ as

$$\mathbf{x}' = \frac{1}{K}\sum_{k=1}^{K} \frac{x^{*0}_{j_k}}{\xi_{j_k}} \cdot \mathbf{e}_{j_k}.$$

With $\mathbf{x}'$, we can bound the objective value of P^{feas}_J as follows:

$$\begin{aligned} v(P^{\text{feas}}_J) &\leq \|\mathbf{A}\mathbf{x}' - \mathbf{b}\|_1 \\ &= \|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}^{*0} + \mathbf{A}\mathbf{x}^{*0} - \mathbf{b}\|_1 \\ &\leq \|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}^{*0}\|_1 + \|\mathbf{A}\mathbf{x}^{*0} - \mathbf{b}\|_1 \\ &= \|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}^{*0}\|_1 + v(P^{\text{feas}}_{\text{distr}}) \end{aligned} \tag{EC.16}$$

where the first step follows by the fact that $\mathbf{x}'$, when restricted to the indices in J, is a feasible solution of P_J^{feas}; the third step follows by the triangle inequality; and the fourth follows by the definition of $\mathbf{x}^{*0}$ as an optimal solution of $P_{\text{distr}}^{\text{feas}}$.

The only remaining step is to bound $\|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}^{*0}\|_1$. To do this, let us define the vector $\mathbf{v}_k$ as

$$\mathbf{v}_k = \frac{x_{j_k}^{*0}}{\xi_{j_k}} \mathbf{A}_{j_k}$$

for each $k \in [K]$. The vectors $\mathbf{v}_1, \ldots, \mathbf{v}_K$ are special for three reasons. First, their sample mean is exactly

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K} \mathbf{v}_k &= \frac{1}{K}\sum_{k=1}^{K} \frac{x_{j_k}^{*0}}{\xi_{j_k}} \mathbf{A}_{j_k} \\
&= \frac{1}{K}\sum_{k=1}^{K} \frac{x_{j_k}^{*0}}{\xi_{j_k}} \mathbf{A}\mathbf{e}_{j_k} \\
&= \mathbf{A}\mathbf{x}'.
\end{aligned}$$

Second, letting $\mathbf{v}$ denote a random variable following the same distribution as each $\mathbf{v}_k$, the expected value of each $\mathbf{v}_k$ is

$$\begin{aligned}
\mathbb{E}[\mathbf{v}] &= \sum_{j \in I_+} \xi_j \cdot \frac{x_j^{*0}}{\xi_j} \mathbf{A}_j \\
&= \sum_{j \in I_+} x_j^{*0} \mathbf{A}_j \\
&= \sum_{j \in [n]} x^{*0_j} \mathbf{A}_j \\
&= \mathbf{A}\mathbf{x}^{*0}
\end{aligned}$$

where I_+ is the subset of indices in $[n]$ such that $\xi_j > 0$. Note that the third step is justified by observing that $\xi_j^{*0} = 0$ whenever $j \notin I_+$ (this is because of the constraint $\mathbf{0} \leq \mathbf{x} \leq C\boldsymbol{\xi}$ in the definition of $P_{\text{distr}}^{\text{feas}}$).

Lastly, observe that each $\mathbf{v}_k$ is bounded as

$$\|\mathbf{v}_k\|_\infty = \frac{x_{j_k}^{*0}}{\xi_{j_k}} \cdot \|\mathbf{A}_{j_k}\|_\infty \leq C \cdot H,$$

where we use the hypothesis that $\|\mathbf{A}_j\|_\infty \leq \|\mathbf{A}\|_{\max}$ and the fact that $\mathbf{x}^{*0}$ satisfies $\mathbf{0} \leq \mathbf{x}^{*0} \leq C\boldsymbol{\xi}$.

With all of these properties, the quantity $\|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}^{*0}\|_1$ can be re-written as $\|(1/K)\sum_{k=1}^{K} \mathbf{v}_k - \mathbb{E}[\mathbf{v}]\|_1$, which we can bound using Lemma EC.2 (see Section B.1). Invoking Lemma EC.2, we get that

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}^{*0}\|_1 &= \|\frac{1}{K}\sum_{k=1}^{K} \mathbf{v}_k - \mathbb{E}[\mathbf{v}]\|_1 \\
&\leq \frac{mC\|\mathbf{A}\|_{\max}}{\sqrt{K}} \left(1 + \sqrt{2\log\frac{1}{\delta}}\right).
\end{aligned}$$

with probability at least $1-\delta$. Using this within the bound (EC.16), we obtain that

$$\begin{aligned} v(P_J^{\text{feas}}) &\leq v(P_{\text{distr}}^{\text{feas}}) + \|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}^{*0}\|_1 \\ &\leq v(P_{\text{distr}}^{\text{feas}}) + \frac{C}{\sqrt{K}} \cdot m \cdot \|\mathbf{A}\|_{\max} \cdot \left(1 + \sqrt{2\log\frac{1}{\delta}}\right) \end{aligned}$$

holds with probability at least $1-\delta$, which completes the proof. □

B.4 Proof of Theorem EC.1 and EC.2

As with Theorems 1 and 2, we first establish analogs of Propositions EC.3 and EC.4 for Algorithm 7.

PROPOSITION EC.5. *Let C be a nonnegative constant and let $P_{J\cup J_F}$ be the column-randomized LP solved by Algorithm 7. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, then*

$$\Delta v(P_{J\cup J_F}) \leq \Delta v(P_{distr}) + \frac{C}{\sqrt{K}} \cdot (1 + \|\mathbf{p}\|_\infty \cdot m \cdot \|\mathbf{A}\|_{\max}) \cdot \left(1 + \sqrt{2\log\frac{2}{\delta}}\right)$$

for any optimal solution $\mathbf{p}$ of problem $D_{J\cup J_F}$.

Proof: The proof of Proposition EC.5 follows along similar lines as the proof of Proposition EC.3. More specifically, we construct $\mathbf{x}'$ and $\mathbf{b}'$ in the same way, and Steps 1 and 2 follow through identically. In the last step, Step 3, the sequence of bounding steps is almost the same, with a few differences:

$$\begin{aligned} v(P_{J\cup J_F}) &= v(P_{J\cup J_F}(\mathbf{b})) \\ &\leq v(P_{J\cup J_F}(\mathbf{b}')) + \mathbf{p}^T(\mathbf{b}-\mathbf{b}') \\ &\leq \mathbf{c}^T\mathbf{x}' + \mathbf{p}^T(\mathbf{b}-\mathbf{b}') \\ &= \mathbf{c}^T\mathbf{x}^{*0} + \mathbf{c}^T(\mathbf{x}'-\mathbf{x}^{*0}) + \mathbf{p}^T(\mathbf{b}-\mathbf{b}') \\ &= v(P_{\text{distr}}) + \mathbf{c}^T(\mathbf{x}'-\mathbf{x}^{*0}) + \mathbf{p}^T(\mathbf{b}-\mathbf{b}') \\ &\leq v(P_{\text{distr}}) + \|\mathbf{c}\|_2\|\mathbf{x}'-\mathbf{x}^{*0}\|_2 + \|\mathbf{p}\|_\infty\|\mathbf{b}-\mathbf{b}'\|_1. \end{aligned}$$

In the above, there are two important, subtle differences in the bounding. First, $\mathbf{p}$ is now any optimal dual solution of $P_{J\cup J_F}$, whereas in Proposition EC.3, we required $\mathbf{p}$ to be any optimal dual solution of P_J. Additionally, the second inequality follows because $\mathbf{x}'$, which we defined as

$$\mathbf{x}' = \frac{1}{K}\sum_{k=1}^{K} \frac{x_{j_k}^{*0}}{\xi_{j_k}} \mathbf{e}_{j_k},$$

is still a feasible solution of $P_{J\cup J_F}(\mathbf{b}')$. (Note that $\mathbf{x}'$ is supported on J, which is obviously a subset of $J\cup J_F$; additionally, $\mathbf{b}'$ was defined as $\mathbf{b}' \equiv \mathbf{A}\mathbf{x}'$, so by construction $\mathbf{x}'$ must satisfy the equality constraint, and by construction $\mathbf{x}'$ is nonnegative, so it satisfies the nonnegativity constraint.)

The remaining steps, which involve applying the high probability bounds from Steps 1 and 2 to $\|\mathbf{x}'-\mathbf{x}^{*0}\|_2$ and $\|\mathbf{b}-\mathbf{b}'\|_1$, follow in the same way as in the proof of Proposition EC.3. □

We now prove Theorem EC.1.

Proof of Theorem EC.1: As in the proof of Theorem 1, we invoke Proposition EC.5 with $\mathbf{p}$ set to an optimal basic feasible solution of $D_{J\cup J_F}$. Note that such a solution exists because Algorithm 7 guarantees that $\text{rank}(\mathbf{A}_{J\cup J_F}) = m$. Since $\mathbf{p}$ is a basic solution of $D_{J\cup J_F}$ it remains a basic solution of the complete dual problem D, and thus it obeys $\|\mathbf{p}\|_\infty \leq \gamma$, which establishes the theorem. □

To establish Theorem EC.2, we similarly need an analog of Proposition EC.4 for Algorithm 7.

PROPOSITION EC.6. *Let C, P_J and P_{distr} be defined as in the statement of Proposition EC.3. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, the following holds: if P_J is feasible, then*

$$\Delta v(P_J) \leq \Delta v(P_{distr}) + \frac{C}{\sqrt{K}} \cdot \|\mathbf{c}^T - \mathbf{p}^T\mathbf{A}\|_2 \cdot \left(1 + \sqrt{2\log\frac{1}{\delta}}\right)$$

for any optimal solution $\mathbf{p}$ of problem D_J (the dual of problem P_J).

Proof: As with Proposition EC.5, we construct $\mathbf{x}'$ and $\mathbf{b}'$ as in the proof of Proposition EC.3, and follow Steps 1 and 2 from that proof. We then follow the bounding procedure in the proof of Proposition EC.4, with some minor modifications:

$$\begin{aligned}
v(P_{J\cup J_F}) &= v(P_{J\cup J_F}(\mathbf{b})) \\
&\leq v(P_{J\cup J_F}(\mathbf{b}')) + \mathbf{p}^T(\mathbf{b}-\mathbf{b}') \\
&\leq \mathbf{c}^T\mathbf{x}' + \mathbf{p}^T(\mathbf{b}-\mathbf{b}') \\
&= \mathbf{c}^T\mathbf{x}^{*0} + \mathbf{c}^T(\mathbf{x}'-\mathbf{x}^{*0}) + \mathbf{p}^T(\mathbf{b}-\mathbf{b}') \\
&= v(P_{\text{distr}}) + \mathbf{c}^T(\mathbf{x}'-\mathbf{x}^{*0}) + \mathbf{p}^T(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{x}') \\
&= v(P_{\text{distr}}) + (\mathbf{c}^T - \mathbf{p}^T\mathbf{A})(\mathbf{x}'-\mathbf{x}^{*0}) \\
&\leq v(P_{\text{distr}}) + \|\mathbf{c}^T - \mathbf{p}^T\mathbf{A}\|_2\|\mathbf{x}'-\mathbf{x}^{*0}\|_2,
\end{aligned}$$

where the main difference from the proof of Proposition EC.4 is again that $\mathbf{p}$ is a dual optimal solution of $P_{J\cup J_F}$, and we use the fact that $\mathbf{x}'$, which is supported on J, is a feasible solution of $P_{J\cup J_F}(\mathbf{b}')$. From here, the rest of the proof is the same as Proposition EC.4. □

We now prove Theorem EC.2.

Proof of Theorem EC.2: As in the proof of Theorem 2, we invoke Proposition EC.6 and set $\mathbf{p}$ to be an optimal basic feasible solution of the dual problem $D_{F\cup F_J}$. Since $\mathbf{p}^T = \mathbf{c}_B^T\mathbf{A}_B^{-1}$ for some set of basic variables $B \subset J\cup J_F \subset [n]$, the dual slack vector $\mathbf{c}^T - \mathbf{p}^T\mathbf{A}$ is the reduced cost vector $\bar{\mathbf{c}}$ of the basis B within the full problem P, and using the assumption that any such $\bar{\mathbf{c}}$ obeys $\|\bar{\mathbf{c}}\|_2 \leq \chi$, the result follows. □

Appendix C Omitted Proofs and Other Results for Section 4

C.1 Proof of Theorem 3

For the solutions $\mathbf{x}^1, \ldots, \mathbf{x}^M$, consider the averaged solution $\tilde{\mathbf{x}}$ defined as

$$\tilde{x}_j = \frac{1}{M} \sum_{i=1}^{M} x_j^i,$$

for each column $j \in [n]$. Since each column j is in at most R of the bases $B^1, \ldots B^M$, any coordinate j of $\tilde{\mathbf{x}}$ is the average of M values of which at most R have non-zero values, and each of those at most R values is upper bounded by $x_{\max}$. It thus follows that for all j,

$$\tilde{x}_j \leq \frac{R}{M} x_{\max}.$$

Observe now that by setting $C = n \cdot (R/M) x_{\max}$, we obtain that

$$C\xi_j = n \cdot (R/M) x_{\max} \cdot 1/n = (R/M) x_{\max},$$

which means that $\tilde{\mathbf{x}}$ satisfies the constraint $x_j \leq C\xi_j$ for all j. Since $\tilde{\mathbf{x}}$ is the convex combination of BFSs to P, it satisfies $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{0}$. Thus $\tilde{\mathbf{x}}$ is a feasible solution to P_{distr}. We therefore have

$$\begin{aligned} v(P_{\text{distr}}) &\leq \mathbf{c}^T \tilde{\mathbf{x}} \\ &= \frac{1}{M} \sum_{i=1}^{M} \mathbf{c}^T \mathbf{x}^i \\ &\leq \frac{1}{M} \sum_{i=1}^{M} (v(P) + \epsilon) \\ &= v(P) + \epsilon, \end{aligned}$$

where the first inequality follows since $\tilde{\mathbf{x}}$ is feasible for P_{distr} and the second inequality follows since each of the M BFSs is assumed to be within ϵ of $v(P)$. Subtracting $v(P)$ from both sides gives the desired result. □

C.2 Additional comments on the generative model 1

We complete our remarks on the generative model 1.

An alternative interpretation An alternative interpretation of generative model 1 can also be obtained in the case when $\eta = 1$ and when $\mathbf{A}$ is structured as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}' \\ \mathbf{1}^T \end{bmatrix},$$

where $\mathbf{1}$ is an n-dimensional vector of ones, and $\mathbf{A}'$ is a $(m-1)$-by-n matrix. In this case, we can see that for any $\boldsymbol{\theta}$ in the $(n-1)$ dimensional simplex, we will have

$$\mathbf{b} = \mathbf{A}\boldsymbol{\theta} = \begin{bmatrix} \mathbf{A}'\boldsymbol{\theta} \\ \mathbf{1}^T\boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \mathbf{A}'\boldsymbol{\theta} \\ 1 \end{bmatrix},$$

which implies that P can be written as

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^T\mathbf{x} && \text{(EC.17a)} \\ \text{subject to} \quad & \mathbf{A}'\mathbf{x} = \mathbf{A}'\boldsymbol{\theta}, && \text{(EC.17b)} \\ & \mathbf{1}^T\mathbf{x} = 1, && \text{(EC.17c)} \\ & \mathbf{x} \geq \mathbf{0}. && \text{(EC.17d)} \end{aligned}$$

We can think of problem (EC.17) as an estimation problem over the space of discrete probability distributions on $[n]$. In particular, constraints (EC.17c) and (EC.17d) enforce that $\mathbf{x}$ is a probability distribution, while constraint (EC.17b) can be interpreted as a constraint that enforces a set of moments of $\mathbf{x}$ to match those of $\boldsymbol{\theta}$. From this perspective, generative model 1 can be loosely interpreted as imposing a uniform prior. The nonparametric choice estimation problem that we numerically study in Section 6 can be regarded as an instance of the moment problem (EC.17) with some modifications.

Scaling c so that $v(P) \geq 0$. A key element of generative model 1 is that $\mathbf{c}$ is selected so that $v(P) \geq 0$. We note that this can always be accomplished: since $\mathbf{x} \geq \mathbf{0}$, any nonnegative choice of $\mathbf{c}$ will ensure that $v(P) \geq 0$, no matter what $\mathbf{A}$ and $\mathbf{b}$ are. In addition, note that the assumption of $v(P) \geq 0$ is actually without loss of generality. If $v(P) < 0$, then let B be an optimal basis, for which the corresponding reduced cost vector $\bar{\mathbf{c}}$ satisfies $\bar{\mathbf{c}} \geq \mathbf{0}$. (Although an optimal BFS may have negative reduced costs due to degeneracy, an optimal BFS and corresponding basis B with a nonnegative reduced cost vector $\bar{\mathbf{c}}$ can be obtained by applying the simplex algorithm with an anticycling pivoting rule such as Bland's rule; see Chapter 3 of Bertsimas and Tsitsiklis 1997.) We can then re-write P as

$$\begin{aligned} & \min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \\ &= \min\{\mathbf{c}_B^T\mathbf{x}_B + \mathbf{c}_N^T\mathbf{x}_N \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \\ &= \min\{\mathbf{c}_B^T\mathbf{A}_B^{-1}(\mathbf{b} - \mathbf{A}_N\mathbf{x}_N) + \mathbf{c}_N^T\mathbf{x}_N \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \\ &= \mathbf{c}_B^T\mathbf{A}_B^{-1}\mathbf{b} + \min\{[\mathbf{c}_N^T - \mathbf{c}_B^T\mathbf{A}_B^{-1}\mathbf{A}_N]\mathbf{x}_N \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \\ &= v(P) + \min\{\bar{\mathbf{c}}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}, \end{aligned}$$

where we observe that the problem $P' \equiv \min\{\bar{\mathbf{c}}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is such that P' and P have the same feasible region and optimal solutions, and $v(P') = 0$. Thus, by replacing $\mathbf{c}$ with the reduced cost vector $\bar{\mathbf{c}}$ we obtain an equivalent problem, up to a constant shift. By further normalizing $\bar{\mathbf{c}}$ to have unit norm, we can ensure that the last step of Algorithm 2 can be accomplished.

C.3 Proof of Theorem 4

In this section, we establish Theorem 4 for generative model 1. The first key result we require is Lemma EC.4, which allows us to bound the gap of the distributional counterpart in terms of an upper bound β on the minimum infinity norm attainable in the polyhedron P.

LEMMA EC.4. *Suppose that P is feasible, $v(P) \geq 0$ and $\beta \geq \min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$. Suppose that $\boldsymbol{\xi}$ is the uniform distribution over $[n]$, i.e., $\xi_j = 1/n$ for all $j \in [n]$. If $C = n\beta$, then P_{distr} is feasible and we have that*

$$\Delta v(P_{distr}) \leq \sqrt{n}\beta.$$

Proof: If $C = n\beta$, then

$$\begin{aligned} P_{\text{distr}} &= \min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{0} \leq \mathbf{x} \leq C\boldsymbol{\xi}\} \\ &= \min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{0} \leq \mathbf{x} \leq n\beta \cdot (1/n) \cdot \mathbf{1}\} \\ &= \min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{0} \leq \mathbf{x} \leq \beta \cdot \mathbf{1}\}, \end{aligned}$$

which must be feasible; this follows by the definition of β as an upper bound on the minimum infinity norm of any feasible solution to P, which itself is assumed to be feasible. Now, observe that for any feasible solution $\mathbf{x}$ of P_{distr}, we have

$$\begin{aligned} \mathbf{c}^T\mathbf{x} &\leq \|\mathbf{c}\|_2 \cdot \|\mathbf{x}\|_2 \\ &= \sqrt{\sum_{j=1}^{n} x_j^2} \\ &\leq \sqrt{\sum_{j=1}^{n} C^2\xi_j^2} \\ &= C\sqrt{\sum_{j=1}^{n} (1/n)^2} \\ &= C/\sqrt{n} \\ &= \sqrt{n}\beta, \end{aligned}$$

where the first inequality follows by Cauchy-Schwartz, and the second inequality by the constraint $\mathbf{x} \leq \beta \cdot \mathbf{1}$. This implies that $v(P_{\text{distr}}) \leq \sqrt{n}\beta$. By the assumption that $v(P) \geq 0$, we thus have that

$$\Delta v(P_{\text{distr}}) = v(P_{\text{distr}}) - v(P) \leq \sqrt{n}\beta - 0 = \sqrt{n}\beta,$$

as required. □

Lemma EC.4 is a general result that is independent of the generative model chosen; we shall use it later when establishing guarantees for generative models 2 and 3.

The next auxiliary result we need is a result on ordered uniform spacings. Let n' be an integer, and suppose that $X_1, \ldots, X_{n'}$ are independent uniformly distributed random variables on $[0,1]$. Define $X_{0,n'} = 0$, $X_{n'+1,n'} = 1$, and define $X_{1,n'}, \ldots, X_{n',n'}$ as the order statistics of $X_1, \ldots, X_{n'}$. Define $\Delta_{k:n'} = X_{k,n'} - X_{k-1,n'}$ for $k = 1, \ldots, n'+1$ as the *(uniform) spacings* of the sample $X_1, \ldots, X_{n'}$. Finally, define the *ordered uniform spacings* $\Delta_{1,n'}, \ldots, \Delta_{n'+1,n'}$ as the order statistics of $\Delta_{1:n'}, \ldots, \Delta_{n'+1:n'}$. The following lemma is a known result on ordered uniform spacings (see Bairamov et al. 2010).

Lemma EC.5. *(Bairamov et al. 2010, Section 3.) For any $k = 1, \ldots, n'+1$,*

$$\mathbb{E}[\Delta_{k,n'}] = \frac{1}{n'+1} \sum_{i=n'+2-k}^{n'+1} \frac{1}{i}.$$

The uniform spacings $(\Delta_{1:n'}, \ldots, \Delta_{n'+1:n'})$ are useful because their joint distribution is uniform on the n'-dimensional unit simplex, which is identical to the Dirichlet$(1, \ldots, 1)$ distribution (see equation 2.1 of Pyke 1965). The ordered uniform spacings are useful because the largest such ordered uniform spacing, $\Delta_{n'+1,n'}$, is exactly the maximum value of a Dirichlet$(1, \ldots, 1)$ random vector. The expected value of this largest ordered uniform spacing will be essential to being able to obtain a high probability bound on the minimum infinity norm solution of P, which is the focus of our next lemma.

Lemma EC.6. *Suppose that P is generated according to generative model 1. Let $t \geq 1$. Then, with probability at least $1 - 1/t$, we have*

$$\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \leq \frac{t\eta(1+\log n)}{n}.$$

Proof: Observe that by the definition of generative model 1, we know that $\mathbf{b} = \mathbf{A}(\eta\boldsymbol{\theta})$ for a $\boldsymbol{\theta}$ drawn from the Dirichlet$(1, \ldots, 1)$ distribution. Since this implies that $\eta\boldsymbol{\theta}$ is a feasible solution of P, we immediately have

$$\begin{aligned}
&\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \\
&\leq \|\eta\boldsymbol{\theta}\|_\infty \\
&= \eta \max_{j \in [n]} \theta_j.
\end{aligned}$$

For the random variable $\max_{j\in[n]}\theta_j$, we can bound its expected value as

$$\begin{aligned}\mathbb{E}[\max_{j\in[n]}\theta_j] &= \mathbb{E}[\Delta_{n,n-1}] \\ &= \frac{1}{n-1+1}\sum_{i=n-1+2-n}^{n-1+1}\frac{1}{i} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{i} \\ &\leq \frac{1+\log n}{n}.\end{aligned}$$

In the above, the steps are as follows. The first step follows because given a sample of $n-1$ i.i.d. uniform random variables, the n unordered spacings $\Delta_{1:n-1},\ldots,\Delta_{n:n-1}$ are distributed in the same way as $\boldsymbol{\theta}$ (i.e., they follow a Dirichlet$(1,\ldots,1)$ distribution). Thus, the nth ordered spacing $\Delta_{n,n-1}$, which is the maximum of $\Delta_{1:n-1},\ldots,\Delta_{n:n-1}$, is distributed the same way as $\max_{j\in[n]}\theta_j$. The second step follows by Lemma EC.5. The third step follows by algebra. The last step follows by using the bound $\sum_{i=2}^{n}\frac{1}{i}\leq\int_1^n\frac{1}{s}ds=\log n$.

Using this bound on the expected value, an application of Markov's inequality implies that with probability at least $1-1/t$,

$$\max_{j\in[n]}\theta_j\leq\frac{t(1+\log n)}{n}.$$

Thus, with probability at least $1-1/t$,

$$\begin{aligned}&\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x}=\mathbf{b},\mathbf{x}\geq\mathbf{0}\} \\ &\leq \eta\max_{j\in[n]}\theta_j \\ &\leq \frac{t\eta(1+\log n)}{n},\end{aligned}$$

as required. □

We now prove Theorem 4.

Proof of Theorem 4: We know that P is feasible, since $\mathbf{x}=\eta\boldsymbol{\theta}$ is a feasible solution, and that $v(P)\geq 0$, which is just by definition of generative model 1. By Lemma EC.6, we have that $\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x}=\mathbf{b},\mathbf{x}\geq\mathbf{0}\}$ is bounded by $\beta=t\eta(1+\log n)/n$ with probability at least $1-1/t$. Therefore, by Lemma EC.4, it follows that when $C=n\beta=t\eta(1+\log n)$, we will have that with probability at least $1-1/t$, that P_{distr} is feasible and the following holds:

$$\begin{aligned}\Delta v(P_{\text{distr}}) &\leq \sqrt{n}\beta \\ &= \sqrt{n}\cdot\frac{t\eta(1+\log n)}{n} \\ &= \frac{t\eta(1+\log n)}{\sqrt{n}},\end{aligned}$$

as required. □

C.4 Proof of Theorem 5

To prove Theorem 5, we begin with two simple results on the behaviors of the random vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n$. As a preview of the later results, we will need to bound the expected value of the supremum of the deviation of the sample average of $(\mathbf{v}^T\mathbf{A}_1)_+, \ldots, (\mathbf{v}^T\mathbf{A}_n)_+$, where $\mathbf{v}$ is an m-dimensional unit norm vector, from its expected value. To do this, we will essentially use the Rademacher complexity of the class of functions of the form $f_{\mathbf{v}}(\tilde{\mathbf{A}}) = \mathbf{v}^T\tilde{\mathbf{A}}$ over all unit norm vectors $\mathbf{v}$. The first result, Lemma EC.7, will allow us to eliminate the $(\cdot)_+$ function when we eventually bound this Rademacher complexity, while the second result, Lemma EC.8, will allows us to bound the simplified expression that results from Lemma EC.7.

LEMMA EC.7. *Suppose that σ_j is a Rademacher variable, i.e., it takes the values -1 and +1 each with probability 1/2; $\mathbf{A}_j$ is a random vector drawn from a standard multivariate normal distribution on $\mathbb{R}^m$; and Y_j is a Bernoulli$(1/2)$ random variable. Suppose that all three random variables are independent. Then, for any vector $\mathbf{v} \in \mathbb{R}^m$, the random variables $\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+$ and $\mathbf{v}^T Y_j \mathbf{A}_j$ have the same distribution.*

Proof: Suppose that $t < 0$. Then we have

$$\begin{aligned}
\mathbf{Pr}(\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+ \leq t) &= \mathbf{Pr}(\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+ \leq t \mid \sigma_j = +1)\mathbf{Pr}(\sigma_j = +1) + \mathbf{Pr}(\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+ \leq t \mid \sigma_j = -1)\mathbf{Pr}(\sigma_j = -1) \\
&= \mathbf{Pr}((\mathbf{v}^T\mathbf{A}_j)_+ \leq t) \cdot (1/2) + \mathbf{Pr}(-(\mathbf{v}^T\mathbf{A}_j)_+ \leq t) \cdot (1/2) \\
&= \mathbf{Pr}((\mathbf{v}^T\mathbf{A}_j)_+ \geq -t) \cdot (1/2) \\
&= \mathbf{Pr}(\mathbf{v}^T\mathbf{A}_j \geq -t) \cdot (1/2) \\
&= \mathbf{Pr}(\mathbf{v}^T\mathbf{A}_j \leq t) \cdot (1/2),
\end{aligned}$$

whereas

$$\begin{aligned}
\mathbf{Pr}(\mathbf{v}^T Y_j \mathbf{A}_j \leq t) &= \mathbf{Pr}(\mathbf{v}^T Y_j \mathbf{A}_j \leq t \mid Y_j = 1)\mathbf{Pr}(Y_j = 1) + \mathbf{Pr}(\mathbf{v}^T Y_j \mathbf{A}_j \leq t \mid Y_j = 0)\mathbf{Pr}(Y_j = 0) \\
&= \mathbf{Pr}(\mathbf{v}^T\mathbf{A}_j \leq t) \cdot (1/2) + 0 \cdot (1/2) \\
&= \mathbf{Pr}(\mathbf{v}^T\mathbf{A}_j \leq t) \cdot (1/2).
\end{aligned}$$

On the other hand, suppose $t \geq 0$. Then we have

$$\begin{aligned}
\mathbf{Pr}(\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+ \leq t) &= \mathbf{Pr}(\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+ \leq t \mid \sigma_j = +1)\mathbf{Pr}(\sigma_j = +1) + \mathbf{Pr}(\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+ \leq t \mid \sigma_j = -1)\mathbf{Pr}(\sigma_j = -1) \\
&= \mathbf{Pr}((\mathbf{v}^T\mathbf{A}_j)_+ \leq t) \cdot (1/2) + \mathbf{Pr}(-(\mathbf{v}^T\mathbf{A}_j)_+ \leq t) \cdot (1/2) \\
&= \mathbf{Pr}(\mathbf{v}^T\mathbf{A}_j \leq t)(1/2) + (1)(1/2),
\end{aligned}$$

whereas

$$\begin{aligned}\mathbf{Pr}(\mathbf{v}^T Y_j \mathbf{A}_j \leq t) &= \mathbf{Pr}(\mathbf{v}^T Y_j \mathbf{A}_j \leq t \mid Y_j = 1)\mathbf{Pr}(Y_j = 1) + \mathbf{Pr}(\mathbf{v}^T Y_j \mathbf{A}_j \leq t \mid Y_j = 0)\mathbf{Pr}(Y_j = 0) \\ &= \mathbf{Pr}(\mathbf{v}^T \mathbf{A}_j \leq t)(1/2) + (1)(1/2),\end{aligned}$$

as desired. □

LEMMA EC.8. *Let $Y_1, \ldots, Y_n$ be sampled independently from a Bernoulli$(1/2)$ distribution and $\mathbf{A}_1, \ldots, \mathbf{A}_n$ be sampled independently from a standard multivariate normal distribution, i.e., $\mathbf{A}_1, \ldots, \mathbf{A}_n \sim$ Normal$(\mathbf{0}, \mathbf{I})$. Then*

$$\mathbb{E}\left\| \frac{1}{n} \sum_{j=1}^{n} Y_j \mathbf{A}_j \right\|_2 \leq \frac{\sqrt{m}}{\sqrt{2n}}.$$

Proof: We have

$$\begin{aligned}&\mathbb{E}\left\| \frac{1}{n} \sum_{j=1}^{n} Y_j \mathbf{A}_j \right\|_2 \\ &\leq \sqrt{\mathbb{E}\left\| \frac{1}{n} \sum_{j=1}^{n} Y_j \mathbf{A}_j \right\|_2^2} \\ &= \sqrt{\mathbb{E}\left[\frac{1}{n^2} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} Y_{j_1} Y_{j_2} \mathbf{A}_{j_1}^T \mathbf{A}_{j_2} \right]} \\ &= \sqrt{\frac{1}{n^2} \sum_{j=1}^{n} \mathbb{E}[Y_j^2] \mathbb{E}[\|\mathbf{A}_j\|_2^2]} \\ &= \sqrt{\frac{1}{n^2} \sum_{j=1}^{n} (1/2) m} \\ &= \sqrt{\frac{n \cdot m}{2n^2}} \\ &= \frac{\sqrt{m}}{\sqrt{2n}},\end{aligned}$$

where the first step follows by Jensen's inequality; the second comes from the definition of the squared norm of a vector as the inner product of that vector with itself; the third comes from the fact that each term $Y_{j_1} Y_{j_2} \mathbf{A}_{j_1}^T \mathbf{A}_{j_2}$ has an expected value of zero when $j_1 \neq j_2$ (since $\mathbf{A}_{j_1}$ and $\mathbf{A}_{j_2}$ both have expected value $\mathbf{0}$ and are independent), and the independence of the Y_j and $\mathbf{A}_j$ variables; the fourth comes from the fact $\mathbb{E}[Y_j^2] = \mathbb{E}[Y_j] = 1/2$, while $\|\mathbf{A}_j\|_2^2$ is a chi-squared random variable with degrees of freedom m, so $\mathbb{E}\|\mathbf{A}_j\|_2^2 = m$; and the fifth and sixth steps follow by algebra. □

The next result we will need is a bound on the aforementioned expected supremum of the deviation of the sample average of $(\mathbf{v}^T\mathbf{A}_1)_+, \ldots, (\mathbf{v}^T\mathbf{A}_n)_+$ from its expected value.

LEMMA EC.9. *Let $\tilde{\mathbf{A}}, \mathbf{A}_1, \ldots, \mathbf{A}_n$ be sampled independently from a standard multivariate normal distribution, i.e., $\tilde{\mathbf{A}}, \mathbf{A}_1, \ldots, \mathbf{A}_n \sim$ Normal$(\mathbf{0}, \mathbf{I})$. Then*

$$\mathbb{E}\left[\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\mathbb{E}(\mathbf{v}^T\tilde{\mathbf{A}})_+ - \frac{1}{n}\sum_{j=1}^n(\mathbf{v}^T\mathbf{A}_j)_+\right|\right] \leq \frac{\sqrt{2m}}{\sqrt{n}}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$.

Proof: To prove this, we will use a classical symmetrization argument from statistical learning theory (see for example Mohri et al. 2018, Theorem 3.1). Such techniques are typically used to bound an expected value of the form $\mathbb{E}[\sup_{f\in\mathcal{F}}(\frac{1}{n}\sum_{j=1}^n f(X_j) - \mathbb{E}[f(X)])]$, where $X_1, \ldots, X_n, X$ are i.i.d. random variables and $\mathcal{F}$ is a class of functions, by the Rademacher complexity of $\mathcal{F}$, which is defined as $R(\mathcal{F}) = \mathbb{E}[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{j=1}^n \sigma_j f(X_j)]$, where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher random variables, that is, random variables that are either $+1$ or -1 with probability $1/2$. (For our purposes, it will not be necessary to formally define the Rademacher complexity, because as we will see, our assumption that $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are standard Gaussian random vectors will allow us to bound it directly.)

We have:

$$\begin{aligned}
&\mathbb{E}_{\{\mathbf{A}_j\}}\left[\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\mathbb{E}_{\tilde{\mathbf{A}}}(\mathbf{v}^T\tilde{\mathbf{A}})_+ - \frac{1}{n}\sum_{j=1}^n(\mathbf{v}^T\mathbf{A}_j)_+\right|\right] \\
&= \mathbb{E}_{\{\mathbf{A}_j\}}\left[\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\mathbb{E}_{\{\tilde{\mathbf{A}}_j\}}[\frac{1}{n}\sum_{j=1}^n(\mathbf{v}^T\tilde{\mathbf{A}}_j)_+] - \frac{1}{n}\sum_{j=1}^n(\mathbf{v}^T\mathbf{A}_j)_+\right|\right] \\
&\leq \mathbb{E}_{\{\mathbf{A}_j\}}\left[\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\mathbb{E}_{\{\tilde{\mathbf{A}}_j\}}\left|\frac{1}{n}\sum_{j=1}^n(\mathbf{v}^T\tilde{\mathbf{A}}_j)_+ - \frac{1}{n}\sum_{j=1}^n(\mathbf{v}^T\mathbf{A}_j)_+\right|\right] \\
&\leq \mathbb{E}_{\{\mathbf{A}_j\},\{\tilde{\mathbf{A}}_j\}}\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\frac{1}{n}\sum_{j=1}^n((\mathbf{v}^T\tilde{\mathbf{A}}_j)_+ - (\mathbf{v}^T\mathbf{A}_j)_+)\right| \\
&= \mathbb{E}_{\{\mathbf{A}_j\},\{\tilde{\mathbf{A}}_j\},\{\sigma_j\}}\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\frac{1}{n}\sum_{j=1}^n\sigma_j((\mathbf{v}^T\tilde{\mathbf{A}}_j)_+ - (\mathbf{v}^T\mathbf{A}_j)_+)\right| \\
&\leq \mathbb{E}_{\{\mathbf{A}_j\},\{\tilde{\mathbf{A}}_j\},\{\sigma_j\}}\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left\{\left|\frac{1}{n}\sum_{j=1}^n\sigma_j(\mathbf{v}^T\tilde{\mathbf{A}}_j)_+\right| + \left|\frac{1}{n}\sum_{j=1}^n\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+\right|\right\} \\
&\leq \mathbb{E}_{\{\tilde{\mathbf{A}}_j\},\{\sigma_j\}}\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\frac{1}{n}\sum_{j=1}^n\sigma_j(\mathbf{v}^T\tilde{\mathbf{A}}_j)_+\right| + \mathbb{E}_{\{\mathbf{A}_j\},\{\sigma_j\}}\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\frac{1}{n}\sum_{j=1}^n\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+\right| \\
&= 2\mathbb{E}_{\{\mathbf{A}_j\},\{\sigma_j\}}\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\frac{1}{n}\sum_{j=1}^n\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+\right|
\end{aligned}$$

$$
\begin{aligned}
&= 2\mathbb{E}_{\{\mathbf{A}_j\},\{Y_j\}} \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \left| \frac{1}{n} \sum_{j=1}^{n} \mathbf{v}^T Y_j \mathbf{A}_j \right| \\
&\leq 2\mathbb{E}_{\{\mathbf{A}_j\},\{Y_j\}} \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \left\{ \|\mathbf{v}\|_2 \cdot \left\| \frac{1}{n} \sum_{j=1}^{n} Y_j \mathbf{A}_j \right\|_2 \right\} \\
&= 2\mathbb{E} \left\| \frac{1}{n} \sum_{j=1}^{n} Y_j \mathbf{A}_j \right\|_2 \\
&\leq 2 \frac{\sqrt{m}}{\sqrt{2n}} \\
&= \frac{\sqrt{2m}}{\sqrt{n}}
\end{aligned}
$$

In the above derivation, the steps are as follows. The first step follows by introducing another i.i.d. sample of columns, $\tilde{\mathbf{A}}_1, \ldots, \tilde{\mathbf{A}}_n$, which follow the same standard multivariate normal distribution as $\mathbf{A}_1, \ldots, \mathbf{A}_n$. The second step follows by Jensen's inequality. The third step follows by elementary properties of sup and expectation. The fourth step follows by observing that the random variables $\mathbf{A}_1, \ldots, \mathbf{A}_n, \tilde{\mathbf{A}}_1, \ldots, \tilde{\mathbf{A}}_n$ are exchangeable, and so multiplying the difference $(|\mathbf{v}^T \tilde{\mathbf{A}}_j| - |\mathbf{v}^T \mathbf{A}_j|)$ by σ_j, which is equally likely to be $+1$ (leaving the term unchanged) or -1 (flipping the difference), will leave the overall expectation unchanged. The fifth follows by the triangle inequality and elementary properties of sup, and the sixth by linearity of expectation. The seventh follows by observing that the two expectations in the prior step are identical.

From here, the remaining steps rely on the properties of the distribution of $\mathbf{A}_1, \ldots, \mathbf{A}_n$. In particular, the eighth step follows by applying Lemma EC.7 to assert that the random variable $\sigma_j(\mathbf{v}^T \mathbf{A}_j)_+$ is identically distributed to $\mathbf{v}^T Y_j \mathbf{A}_j$, allowing us to replace the former random variable with the latter random variable and leave the expectation unchanged. The ninth step follows by the Cauchy-Schwartz inequality, and the tenth step by the fact that each $\mathbf{v}$ is unit norm. The tenth and eleventh steps follows by applying Lemma EC.8 and algebra. □

The last auxiliary result we will need is to characterize in closed form the expected value of $\mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+$, where $\tilde{\mathbf{A}}$ is a standard normal random vector (i.e., a column of the matrix $\mathbf{A}$).

LEMMA EC.10. *Suppose that* $\tilde{\mathbf{A}} \sim$ *Normal*$(\mathbf{0}, \mathbf{I})$. *Then for any vector* $\mathbf{p} \in \mathbb{R}^m$,

$$\mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ = \frac{\sqrt{2}}{2\sqrt{\pi}} \|\mathbf{p}\|_2.$$

Proof: We have

$$\mathbb{E}[(\mathbf{p}^T \tilde{\mathbf{A}})_+] = \mathbb{E}[(\mathbf{p}^T \tilde{\mathbf{A}})_+ \mid \mathbf{p}^T \tilde{\mathbf{A}} < 0] \mathbf{Pr}(\mathbf{p}^T \tilde{\mathbf{A}} < 0) + \mathbb{E}[(\mathbf{p}^T \tilde{\mathbf{A}})_+ \mid \mathbf{p}^T \tilde{\mathbf{A}} \geq 0] \mathbf{Pr}(\mathbf{p}^T \tilde{\mathbf{A}} \geq 0)$$

$$
\begin{aligned}
&= (0) \cdot (1/2) + \mathbb{E}[(\mathbf{p}^T \tilde{\mathbf{A}})_+ \mid \mathbf{p}^T \tilde{\mathbf{A}} \geq 0] \cdot (1/2) \\
&= \frac{\sqrt{2}}{\sqrt{\pi}} \sqrt{\mathbf{p}^T \mathbf{I} \mathbf{p}} \cdot (1/2) \\
&= \frac{\sqrt{2}}{2\sqrt{\pi}} \|\mathbf{p}\|_2,
\end{aligned}
$$

where the first step follows by conditioning; the second step follows by the fact that $\mathbf{p}^T \tilde{\mathbf{A}}$ follows a normal distribution with mean 0, and that $(\mathbf{p}^T \tilde{\mathbf{A}})_+ = 0$ when $\mathbf{p}^T \tilde{\mathbf{A}} < 0$; the third step follows by recognizing that the random variable $\left((\mathbf{p}^T \tilde{\mathbf{A}})_+ \mid \mathbf{p}^T \tilde{\mathbf{A}} \geq 0 \right)$ follows the same distribution as $|\mathbf{p}^T \tilde{\mathbf{A}}|$, and $|\mathbf{p}^T \tilde{\mathbf{A}}|$ follows a *half-normal* distribution, whose mean is $\sigma\sqrt{2}/\sqrt{\pi}$, where σ is the standard deviation of $\mathbf{p}^T \tilde{\mathbf{A}}$; and the final step follows by algebra. □

With these auxiliary results in hand, we can now establish the following major result, which provides a high probability bound on the minimum infinity norm of any feasible solution of P.

THEOREM EC.3. *Suppose that P is generated according to generative model 2. Let $t \geq 1$ and suppose that $n > 4\pi t^2 m$. With probability at least $1 - 1/t$, we have that P is feasible and that*

$$
\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} < \frac{\|\mathbf{b}\|_2}{n} \cdot \frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}} - \frac{t\sqrt{2m}}{\sqrt{n}}}.
$$

Proof: Let $\tilde{\mathbf{A}}$ be a random vector that follows the same distribution as $\mathbf{A}_1, \ldots, \mathbf{A}_n$. We have that

$$
\begin{aligned}
&\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \\
&= \max\left\{ \mathbf{p}^T \mathbf{b} \mid \sum_{j=1}^n (\mathbf{p}^T \mathbf{A}_j)_+ \leq 1 \right\} \\
&= \max\left\{ \mathbf{p}^T \mathbf{b} \mid \frac{1}{n} \sum_{j=1}^n (\mathbf{p}^T \mathbf{A}_j)_+ \leq \frac{1}{n} \right\} \\
&= \max\left\{ \mathbf{p}^T \mathbf{b} \mid \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ - \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ + \frac{1}{n} \sum_{j=1}^n (\mathbf{p}^T \mathbf{A}_j)_+ \leq \frac{1}{n} \right\} \\
&= \max\left\{ \mathbf{p}^T \mathbf{b} \mid \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ \leq \frac{1}{n} + \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ - \frac{1}{n} \sum_{j=1}^n (\mathbf{p}^T \mathbf{A}_j)_+ \right\} \\
&\leq \max\left\{ \mathbf{p}^T \mathbf{b} \mid \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ \leq \frac{1}{n} + \|\mathbf{p}\|_2 \cdot \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \left| \mathbb{E}(\mathbf{v}^T \tilde{\mathbf{A}})_+ - \frac{1}{n} \sum_{j=1}^n (\mathbf{v}^T \mathbf{A}_j)_+ \right| \right\} \quad (*)
\end{aligned}
$$

where the first step follows by strong duality; the second, third and fourth step by algebra; and the sixth step by recognizing that

$$
\mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ - \frac{1}{n} \sum_{j=1}^n (\mathbf{p}^T \mathbf{A}_j)_+
$$

$$\leq \|\mathbf{p}\|_2 \cdot \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \left| \mathbb{E}(\mathbf{v}^T \tilde{\mathbf{A}})_+ - \frac{1}{n}\sum_{j=1}^{n} (\mathbf{v}^T \mathbf{A}_j)_+ \right|$$

holds trivially when $\mathbf{p} = \mathbf{0}$, whereas when $\mathbf{p} \neq \mathbf{0}$, we have

$$\begin{aligned}
&\mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ - \frac{1}{n}\sum_{j=1}^{n} (\mathbf{p}^T \mathbf{A}_j)_+ \\
&= \|\mathbf{p}\|_2 \cdot \left[\mathbb{E}\left(\left(\frac{\mathbf{p}}{\|\mathbf{p}\|_2} \right)^T \tilde{\mathbf{A}} \right)_+ - \frac{1}{n}\sum_{j=1}^{n} \left(\left(\frac{\mathbf{p}}{\|\mathbf{p}\|_2} \right)^T \mathbf{A}_j \right)_+ \right] \\
&\leq \|\mathbf{p}\|_2 \cdot \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \left| \mathbb{E}(\mathbf{v}^T \tilde{\mathbf{A}})_+ - \frac{1}{n}\sum_{j=1}^{n} (\mathbf{v}^T \mathbf{A}_j)_+ \right|,
\end{aligned}$$

where the equality follows because $(\cdot)_+$ is positively homogenous, and the inequality follows because $\mathbf{p}/\|\mathbf{p}\|_2$ is a unit norm vector.

To proceed from here, we will now use Lemma EC.9. Recall by Lemma EC.9 that

$$\mathbb{E} \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \left| \mathbb{E}(\mathbf{v}^T \tilde{\mathbf{A}})_+ - \frac{1}{n}\sum_{j=1}^{n} (\mathbf{v}^T \mathbf{A}_j)_+ \right| \leq \frac{\sqrt{2m}}{\sqrt{n}}.$$

Consider the event E defined as

$$E = \left\{ \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} |\mathbb{E}(\mathbf{v}^T \tilde{\mathbf{A}})_+ - \frac{1}{n}\sum_{j=1}^{n} (\mathbf{v}^T \mathbf{A}_j)_+| \leq \frac{t\sqrt{2m}}{\sqrt{n}} \right\}.$$

By Markov's inequality, we have that $\mathbf{Pr}(E) \geq 1 - 1/t$. Thus, with probability $1 - 1/t$, $(*)$ is bounded from above as

$$(*) \leq \max\{\mathbf{p}^T \mathbf{b} \mid \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ \leq \frac{1}{n} + \|\mathbf{p}\|_2 \cdot \frac{t\sqrt{2m}}{\sqrt{n}}\}.$$

We now have

$$\begin{aligned}
&\max\{\mathbf{p}^T \mathbf{b} \mid \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ \leq \frac{1}{n} + \|\mathbf{p}\|_2 \cdot \frac{t\sqrt{2m}}{\sqrt{n}}\} \\
&= \max\{\mathbf{p}^T \mathbf{b} \mid \frac{\sqrt{2}}{2\sqrt{\pi}} \cdot \|\mathbf{p}\|_2 \leq \frac{1}{n} + \|\mathbf{p}\|_2 \cdot \frac{t\sqrt{2m}}{\sqrt{n}}\} \\
&= \max\{\mathbf{p}^T \mathbf{b} \mid \left(\frac{\sqrt{2}}{2\sqrt{\pi}} - \frac{t\sqrt{2m}}{\sqrt{n}} \right) \cdot \|\mathbf{p}\|_2 \leq \frac{1}{n}\} \\
&= \max\{\mathbf{p}^T \mathbf{b} \mid \|\mathbf{p}\|_2 \leq \frac{1}{n} \cdot \frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}} - \frac{t\sqrt{2m}}{\sqrt{n}}}\} \\
&= \frac{\|\mathbf{b}\|_2}{n} \cdot \frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}} - \frac{t\sqrt{2m}}{\sqrt{n}}}
\end{aligned}$$

where the first step follows by the closed form expression for $\mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+$ from Lemma EC.10; the second step follows by algebra; the third step follows by algebra and also by our assumption on n;

and the final step by the fact that $\max\{\mathbf{d}^T\mathbf{x} \mid \|\mathbf{x}\|_2 \leq r\} = r\|\mathbf{d}\|_2$. Note that in the third step, we are using the hypothesis that $n > 4\pi t^2 m$ to ensure that the coefficient $\left(\frac{\sqrt{2}}{2\sqrt{\pi}} - \frac{t\sqrt{2m}}{\sqrt{n}}\right)$ is positive, and that the direction of the inequality in the constraint is unchanged.

We thus have, that with probability at least $1 - 1/t$, that

$$\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \leq \frac{\|\mathbf{b}\|_2}{n} \cdot \frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}} - \frac{t\sqrt{2m}}{\sqrt{n}}},$$

which establishes the required bound on the infinity norm.

To see why P must be feasible, that is, why $\{\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is non-empty, observe that P is feasible if and only if $\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is feasible. Observe that this latter problem is feasible if and only if its dual problem $\max\{\mathbf{p}^T\mathbf{b} \mid \sum_{j=1}^n (\mathbf{p}^T\mathbf{A}_j)_+ \leq 1\}$, which is always feasible, is bounded. By our reasoning above, this problem is bounded with probability at least $1 - 1/t$, and therefore $P \equiv \min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is feasible with probability at least $1 - 1/t$, as required. □

We can now prove Theorem 5.

Proof of Theorem 5: Let β be defined as

$$\beta = \frac{\|\mathbf{b}\|_2}{n} \cdot \frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}} - \frac{t\sqrt{2m}}{\sqrt{n}}}.$$

Observe that the given C in the statement of Theorem 5 is exactly $C = n\beta$.

By Theorem EC.3, with probability at least $1 - 1/t$, it follows that P is feasible, and that

$$\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \leq \beta.$$

Recall by the definition of generative model 2 that whenever P is feasible, we set $\mathbf{c}$ so that $v(P) \geq 0$. Thus, with probability at least $1 - 1/t$, by Lemma EC.4, it follows that P_{distr} is feasible and that

$$\begin{aligned}
\Delta v(P_{\text{distr}}) &\leq \frac{C}{\sqrt{n}} \\
&= \sqrt{n}\beta \\
&= \frac{\|\mathbf{b}\|_2}{\sqrt{n}} \cdot \frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}} - \frac{t\sqrt{2m}}{\sqrt{n}}},
\end{aligned}$$

exactly as required. □

C.5 Generative model 2-U: uniform distribution on the unit sphere

In this section, we discuss an alternate generative model, generative model 2-U, which is closely related to generative model 2. In this new generative model, the columns $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are independently randomly generated from the uniform distribution on the unit sphere $S^m = \{\mathbf{v} \in \mathbb{R}^m \mid \|\mathbf{v}\|_2 = 1\}$ in $\mathbb{R}^m$. Upon generating the columns, we then fix $\mathbf{b}$ and $\mathbf{c}$.

Algorithm 8 Generative Model 2-U

1: Generate n i.i.d. random vectors $\mathbf{A}_1, \ldots, \mathbf{A}_n \sim \text{Uniform}(S^m)$, where $\text{Uniform}(S^m)$ denotes the uniform distribution on the unit sphere $S^m = \{\mathbf{v} \in \mathbb{R}^m \mid \|\mathbf{v}\|_2 = 1\}$.
2: Set $\mathbf{A} = [\mathbf{A}_1 \ \mathbf{A}_2 \ \cdots \mathbf{A}_n]$.
3: Fix any right-hand side vector $\mathbf{b} \in \mathbb{R}^m$.
4: If $\{\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is nonempty, fix any $\mathbf{c} \in \{\mathbf{v} \in \mathbb{R}^n \mid \|\mathbf{v}\|_2 = 1\}$ such that $v(P) \geq 0$; otherwise, fix any $\mathbf{c} \in \{\mathbf{v} \in \mathbb{R}^n \mid \|\mathbf{v}\|_2 = 1\}$.
5: **return** $(\mathbf{A}, \mathbf{b}, \mathbf{c})$.

This model is closely related to generative model 2, because for any standard normal random vector $\mathbf{A}'$, the random vector $\mathbf{A}'/\|\mathbf{A}'\|_2$ is uniformly distributed on the unit sphere S^m. Additionally, as mentioned in Section 4.2, this model is universal, in the sense that any LP of the form $\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ with non-zero columns can be transformed into an equivalent LP where all the columns have unit norm.

The main theoretical result of this section is Theorem EC.4, which asserts that with high probability, the distributional counterpart gap under this generative model is $O(1/\sqrt{n})$. In the statement of the theorem below, Γ is the gamma function, i.e., $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$.

THEOREM EC.4. *Suppose that P is generated according to generative model 2-U. Assume that $\boldsymbol{\xi}$ is the uniform distribution over $[n]$, that is, $\xi_j = 1/n$ for all $j \in [n]$. Let $t \geq 1$, and suppose that $n > 4\pi t^2 \mu_m^2$, where $\mu_m = \sqrt{2} \cdot \Gamma((m+1)/2)/\Gamma(m/2)$. Suppose that C is set as*

$$C = \|\mathbf{b}\|_2 \cdot \frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}\mu_m} - \frac{t\sqrt{2}}{\sqrt{n}}}.$$

Then, with probability at least $1 - 1/t$, we have that P and P_{distr} are feasible, and

$$\Delta v(P_{distr}) \leq \frac{\|\mathbf{b}\|_2}{\sqrt{n}} \cdot \frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}\mu_m} - \frac{t\sqrt{2}}{\sqrt{n}}}.$$

This result is very similar to Theorem 5. The main difference is in the factor which multiplies $\|\mathbf{b}\|_2$ in the particular choice of C and $\|\mathbf{b}\|_2/\sqrt{n}$ in the upper bound on $\Delta v(P_{\text{distr}})$. In the standard Gaussian case, this factor is $[\sqrt{2}/(2\sqrt{\pi}) - t\sqrt{2m}/\sqrt{n}]^{-1}$, whereas in the case of the uniform distribution on the unit sphere, the factor is $[\sqrt{2}/(2\sqrt{\pi}\mu_m) - t\sqrt{2}/\sqrt{n}]^{-1}$. Note that μ_m appears because this is the mean of a chi distributed random variable with m degrees of freedom; this distribution, in turn, appears because this is the distribution of the norm of a standard normal random vector. By Jensen's inequality, μ_m is lower than $\sqrt{m}$, which is the square root of the mean of a chi-squared distributed random variable with m degrees of freedom, but numerically μ_m is actually very close

to $\sqrt{m}$. Thus, comparing the two factors, the factor for generative model 2-U is roughly $\sqrt{m}$ larger. This makes sense, because in generative model 2, the columns will have norm that is on average larger by a factor of $\mu_m \approx \sqrt{m}$ than the columns in generative model 2-U, so the decision variable vector $\mathbf{x}$ should be correspondingly scaled by μ_m to ensure $\mathbf{A}\mathbf{x} = \mathbf{b}$.

We now turn our attention to proving Theorem EC.4. As in the case of generative model 2, we will require a number of auxiliary results. Our first such auxiliary result is an analog of Lemma EC.7, which is a technical result needed to bound the expected supremum of the deviation of the sample average of $(\mathbf{p}^T\mathbf{A}_j)_+$ from its expected value. The proof of this result follows along very similar lines to the proof of Lemma EC.7. The key is that like in the case where $\mathbf{A}_j$ is a standard normal random vector, when $\mathbf{A}_j$ is uniformly distributed on the unit sphere, the distribution of $\mathbf{v}^T\mathbf{A}_j$ is symmetric about zero. For brevity, we omit the proof.

LEMMA EC.11. *Suppose that $\mathbf{A}_j \sim$ Uniform(S^m); suppose that σ_j is a Rademacher random variable (i.e., σ_j is either +1 or -1, both with probability 1/2); and suppose that $Y_j \sim$ Bernoulli$(1/2)$. Suppose that $\mathbf{A}_j$, Y_j and σ_j are independent. Then for any vector $\mathbf{v} \in \mathbb{R}^m$, the random variables $\sigma_j(\mathbf{v}^T\mathbf{A}_j)_+$ and $\mathbf{v}^T Y_j \mathbf{A}_j$ follow the same distribution.*

We next have an analog of Lemma EC.8, which bounds the expected Euclidean norm of the sample average of a particular collection of i.i.d. random vectors. The proof of this lemma is omitted as it follows along essentially the same lines as the proof of Lemma EC.8.

LEMMA EC.12. *Suppose that $\mathbf{A}_1, \ldots, \mathbf{A}_n \sim$ Uniform(S^m) are independent random variables and that $Y_1, \ldots, Y_n \sim$ Bernoulli$(1/2)$ are independent random variables. Then we have*

$$\mathbb{E}\left\| \frac{1}{n}\sum_{j=1}^{n} Y_j \mathbf{A}_j \right\|_2 \leq \frac{1}{\sqrt{2n}}.$$

With Lemma EC.11 and EC.12 in hand, we can prove the following lemma, which is an analog of Lemma EC.9. This lemma allows us to bound the expected supremum of the deviation of the sample average of $(\mathbf{v}^T\mathbf{A}_1)_+, \ldots, (\mathbf{v}^T\mathbf{A}_n)_+$ from its expected value, over all unit vectors $\mathbf{v}$.

LEMMA EC.13. *Suppose that $\tilde{\mathbf{A}}, \mathbf{A}_1, \ldots, \mathbf{A}_n \sim$ Uniform(S^m) are independent random variables distributed uniformly on the unit sphere S^m. Then*

$$\mathbb{E}\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\frac{1}{n}\sum_{j=1}^{n}(\mathbf{p}^T\mathbf{A}_j)_+ - \mathbb{E}(\mathbf{p}^T\tilde{\mathbf{A}})_+\right| \leq \frac{\sqrt{2}}{\sqrt{n}}$$

Proof: The proof follows essentially the same initial steps as that of Lemma EC.9. Following those steps, we obtain the bound

$$\begin{aligned}
&\mathbb{E} \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \left| \frac{1}{n} \sum_{j=1}^{n} (\mathbf{p}^T \mathbf{A}_j)_+ - \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ \right| \\
&\leq 2\mathbb{E} \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \left| \frac{1}{n} \sum_{j=1}^{n} \sigma_j (\mathbf{v}^T \mathbf{A}_j)_+ \right| \\
&= 2\mathbb{E} \sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \left| \frac{1}{n} \sum_{j=1}^{n} \mathbf{v}^T Y_j \mathbf{A}_j \right| \\
&\leq 2\mathbb{E} \left[\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1} \|\mathbf{v}\|_2 \left\| \frac{1}{n} \sum_{j=1}^{n} Y_j \mathbf{A}_j \right\|_2 \right] \\
&\leq 2\mathbb{E} \left\| \frac{1}{n} \sum_{j=1}^{n} Y_j \mathbf{A}_j \right\|_2 \\
&\leq 2 \cdot \frac{1}{\sqrt{2n}} \\
&= \frac{\sqrt{2}}{\sqrt{n}}
\end{aligned}$$

where the first equality follows by Lemma EC.11; the second inequality by Cauchy-Schwartz; the third inequality by the fact that each $\mathbf{v}$ in the sup is of unit norm; and the fourth inequality by Lemma EC.12. □

The last auxiliary result we will need is the following lemma, which provides the closed form expression for $\mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+$ when $\tilde{\mathbf{A}}$ is uniformly distributed on the unit sphere.

LEMMA EC.14. *Suppose that $\tilde{\mathbf{A}} \sim$ Uniform(S^m). Then for any vector $\mathbf{p} \in \mathbb{R}^m$,*

$$\mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ = \frac{\sqrt{2}}{2\sqrt{\pi}\mu_m} \cdot \|\mathbf{p}\|_2.$$

where $\mu_m = \sqrt{2}\Gamma((m+1)/2)/\Gamma(m/2)$, where $\Gamma(\cdot)$ is the gamma function.

Proof: Let $\mathbf{Z} \sim \text{Normal}(\mathbf{0}, \mathbf{I})$ be a standard normal random vector in $\mathbb{R}^m$. Let θ be a random variable that follows the chi distribution with m degrees of freedom, and suppose that θ is independent of $\tilde{\mathbf{A}}$. Then $\mathbf{Z}$ and $\theta\tilde{\mathbf{A}}$ have the same distribution. We therefore have

$$\begin{aligned}
\mathbb{E}(\mathbf{p}^T \mathbf{Z})_+ &= \mathbb{E}(\mathbf{p}^T \theta \tilde{\mathbf{A}})_+ \\
&= \mathbb{E}[\theta \cdot (\mathbf{p}^T \tilde{\mathbf{A}})_+] \\
&= \mathbb{E}[\theta] \cdot \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+ \\
&= \mu_m \mathbb{E}(\mathbf{p}^T \tilde{\mathbf{A}})_+
\end{aligned}$$

where the first step follows by the distributional equivalence of $\mathbf{Z}$ and $\theta\tilde{\mathbf{A}}$; the second by the fact that the $(\cdot)_+$ function is positively homogeneous; the third by the independence of θ and $\tilde{\mathbf{A}}$; and the fourth by the fact that μ_m is precisely the mean of a chi-distributed random variable. Using the fact (Lemma EC.10) that

$$\mathbb{E}(\mathbf{p}^T\tilde{\mathbf{Z}}_+) = \frac{\sqrt{2}}{2\sqrt{\pi}}\|\mathbf{p}\|_2,$$

we obtain that

$$\begin{aligned}\mathbb{E}(\mathbf{p}^T\tilde{\mathbf{A}})_+ &= \frac{1}{\mu_m}\cdot\mathbb{E}(\mathbf{p}^T\mathbf{Z})_+ \\ &= \frac{\sqrt{2}}{2\sqrt{\pi}\mu_m}\|\mathbf{p}\|_2,\end{aligned}$$

as required. □

We are now in a position to prove Theorem EC.5, which is an analog of Theorem EC.3.

THEOREM EC.5. *Suppose that P is generated according to generative model 2-U. Let $t \geq 1$. Suppose that $n > 4t^2\pi\mu_m^2$, where $\mu_m = \sqrt{2}\Gamma((m+1)/2)/\Gamma(m/2)$ is the mean of a chi-distributed random variable with m degrees of freedom. Then, with probability at least $1-1/t$, we have that P is feasible and*

$$\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \leq \frac{\|\mathbf{b}\|_2}{n}\cdot\frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}\mu_m} - \frac{t\sqrt{2}}{\sqrt{n}}}.$$

Proof: Following similar steps as in the proof of Theorem EC.3, we have

$$\begin{aligned}&\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \\ &= \max\left\{\mathbf{p}^T\mathbf{b} \mid \sum_{j=1}^n(\mathbf{p}^T\mathbf{A}_j)_+ \leq 1\right\} \\ &= \max\left\{\mathbf{p}^T\mathbf{b} \mid \mathbb{E}(\mathbf{p}^T\tilde{\mathbf{A}})_+ \leq \frac{1}{n} + \mathbb{E}(\mathbf{p}^T\tilde{\mathbf{A}})_+ - \frac{1}{n}\sum_{j=1}^n(\mathbf{p}^T\mathbf{A}_j)_+\right\} \\ &\leq \max\left\{\mathbf{p}^T\mathbf{b} \mid \mathbb{E}(\mathbf{p}^T\tilde{\mathbf{A}})_+ \leq \frac{1}{n} + \|\mathbf{p}\|_2\cdot\sup_{\mathbf{v}:\|\mathbf{v}\|_2=1}\left|\mathbb{E}(\mathbf{v}^T\tilde{\mathbf{A}})_+ - \frac{1}{n}\sum_{j=1}^n(\mathbf{v}^T\mathbf{A}_j)_+\right|\right\}. \quad (*)\end{aligned}$$

From here, by using Lemma EC.13 in combination with Markov's inequality, we have with probability at least $1-1/t$ that

$$\begin{aligned}(*) &\leq \max\left\{\mathbf{p}^T\mathbf{b} \mid \mathbb{E}(\mathbf{p}^T\tilde{\mathbf{A}})_+ \leq \frac{1}{n} + \|\mathbf{p}\|_2\cdot\frac{t\sqrt{2}}{\sqrt{n}}\right\} \\ &= \max\left\{\mathbf{p}^T\mathbf{b} \mid \frac{\sqrt{2}}{2\sqrt{\pi}\mu_m}\|\mathbf{p}\|_2 \leq \frac{1}{n} + \|\mathbf{p}\|_2\cdot\frac{t\sqrt{2}}{\sqrt{n}}\right\} \\ &= \max\left\{\mathbf{p}^T\mathbf{b} \mid \|\mathbf{p}\|_2 \leq \frac{1}{n}\cdot\frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}\mu_m} - \frac{t\sqrt{2}}{\sqrt{n}}}\right\}\end{aligned}$$

$$= \frac{\|\mathbf{b}\|_2}{n} \cdot \frac{1}{\frac{\sqrt{2}}{2\sqrt{\pi}\mu_m} - \frac{t\sqrt{2}}{\sqrt{n}}},$$

where the first equality follows by applying Lemma EC.14. With regard to the feasibility of P, this again follows by the fact that the dual of $\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$, which is always feasible, is bounded with probability at least $1 - 1/t$. □

C.6 Proof of Theorem 6

To establish Theorem 6, we first require a simple adaptation of Lemma EC.4. The proof is straightforward, and omitted for brevity.

LEMMA EC.15. *Suppose that $P^{covering}$ is feasible, $v(P) \geq 0$ and $\beta \geq \min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$. Suppose that $\boldsymbol{\xi}$ is the uniform distribution over $[n]$, i.e., $\xi_j = 1/n$ for all $j \in [n]$. If $C = n\beta$, then $P^{covering}_{distr}$ is feasible and we have that*

$$\Delta v(P^{covering}_{distr}) \leq \sqrt{n}\beta.$$

We next require the following lemma, which is a concentration result for the minimum of a collection of independent binomial random variables.

LEMMA EC.16. *Let $Y_1, \ldots, Y_m$ be independent random variables, with each $Y_i \sim Binomial(n, q_i)$. Let $\delta \in (0,1)$. Then with probability at least $1 - \delta$, we have*

$$\min_{i \in [m]} Y_i \geq n \cdot \left(\min_{i \in [m]} q_i - \sqrt{\frac{1}{2n} \log \frac{m}{\delta}} \right).$$

Proof: Let $\epsilon > 0$. Then for any $i \in [m]$,

$$\begin{aligned}
&\mathbf{Pr}(Y_i < \min_{i' \in [m]} \mathbb{E}[Y_{i'}] - n\epsilon) \\
&\leq \mathbf{Pr}(Y_i < \mathbb{E}[Y_i] - n\epsilon) \\
&\leq \exp(-\frac{2n^2\epsilon^2}{\sum_{j=1}^n (1-0)^2}) \\
&= \exp(-\frac{2n^2\epsilon^2}{n}) \\
&= \exp(-2n\epsilon^2).
\end{aligned}$$

where the second inequality follows by Hoeffding's inequality.

Now, observe that

$$\begin{aligned}
&\mathbf{Pr}(\min_{i\in[m]} Y_i < \min_{i'\in[m]} \mathbb{E}[Y_{i'}] - n\epsilon) \\
&= \mathbf{Pr}\left(\bigcup_{i\in[m]} \{Y_i < \min_{i'\in[m]} \mathbb{E}[Y_{i'}] - n\epsilon\}\right) \\
&\leq \sum_{i=1}^{m} \mathbf{Pr}(Y_i < \min_{i'\in[m]} \mathbb{E}[Y_{i'}] - n\epsilon) \\
&\leq m \cdot \exp(-2n\epsilon^2),
\end{aligned}$$

where the first inequality follows by the union bound. This implies that

$$\mathbf{Pr}\left(\min_{i\in[m]} Y_i \geq \min_{i'\in[m]} \mathbb{E}[Y_{i'}] - n\epsilon\right) \geq 1 - m \cdot \exp(-2n\epsilon^2).$$

Note that ϵ was arbitrary; to make the right hand side of the previous bound equal to $1-\delta$, we can solve for ϵ as

$$\begin{aligned}
\delta &= m \cdot \exp(-2n\epsilon^2) \\
\Rightarrow \log\delta &= \log m - 2n\epsilon^2 \\
\Rightarrow \epsilon^2 &= \frac{1}{2n}\log\frac{m}{\delta} \\
\Rightarrow \epsilon &= \sqrt{\frac{1}{2n}\log\frac{m}{\delta}}.
\end{aligned}$$

Thus, with probability at least $1-\delta$, we have that

$$\min_{i\in[m]} Y_i \geq \min_{i'\in[m]} \mathbb{E}[Y_{i'}] - \sqrt{\frac{n}{2}\log\frac{m}{\delta}}$$

and noting that $\mathbb{E}[Y_{i'}] = nq_{i'}$, this is equivalent to

$$\min_{i\in[m]} Y_i \geq n \min_{i'\in[m]} q_{i'} - \sqrt{\frac{n}{2}\log\frac{m}{\delta}},$$

as required. □

With this concentration result in hand, we can now prove Theorem EC.6, which states that the minimum infinity norm of any feasible solution of P is $O(1/n)$ with high probability.

THEOREM EC.6. *Suppose that P is generated according to generative model 3. Let $\delta \in (0,1)$. Suppose that $n > \log(m/\delta)/[2(\min_{i\in[m]} q_i)^2]$. Then with probability at least $1-\delta$, $P^{covering}$ is feasible and*

$$\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \leq \frac{\max_{i\in[m]} b_i}{n} \cdot \frac{1}{\min_{i'\in[m]} q_{i'} - \sqrt{\frac{1}{2n}\log\frac{m}{\delta}}}$$

Proof: We have

$$\begin{aligned}
&\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \\
&= \max\{\mathbf{p}^T\mathbf{b} \mid \sum_{j=1}^{n} (\mathbf{p}^T\mathbf{A}_j)_+ \leq 1, \mathbf{p} \geq \mathbf{0}\} \\
&= \max\{\mathbf{p}^T\mathbf{b} \mid \sum_{j=1}^{n} \mathbf{p}^T\mathbf{A}_j \leq 1, \mathbf{p} \geq \mathbf{0}\} \\
&= \max\{\mathbf{p}^T\mathbf{b} \mid \sum_{i=1}^{m} p_i Y_i \leq 1, \mathbf{p} \geq \mathbf{0}\}, \qquad \text{(EC.18)}
\end{aligned}$$

where $Y_i = \sum_{j=1}^{n} A_{i,j}$ for each $i \in [m]$. In the first step, we have simply taken the dual of the original problem; in the second step, we use the fact that $\mathbf{p} \geq \mathbf{0}$ and $\mathbf{A} \geq \mathbf{0}$ to assert that $(\mathbf{p}^T\mathbf{A}_j)_+ = \mathbf{p}^T\mathbf{A}_j$; and in the third, we use the definition of the Y_i's.

By the definition of the generative model, we have that each $Y_i \sim \text{Binomial}(n, q_i)$. Therefore, by Lemma EC.16, we have that with probability at least $1 - \delta$,

$$\min_{i \in [m]} Y_i \geq n \min_{i' \in [m]} q_{i'} - \sqrt{\frac{n}{2} \log \frac{m}{\delta}}. \qquad \text{(EC.19)}$$

In addition, by the assumption on n, it follows that the right hand side of (EC.19) is positive, which implies that $Y_i > 0$ for all $i \in [m]$. Thus, when (EC.19) holds, we can determine the optimal solution of problem (EC.18) as follows: the optimal solution is given by $p_{i^*} = 1/Y_{i^*}$ for $i^* = \arg\max_{i \in [m]} b_i/Y_i$ and $p_i = 0$ for all $i \neq i^*$. (We remind the reader here that the definition of generative model 3 requires $\mathbf{b}$ to be nonnegative. We also note in the case that the arg max is not a singleton, we can set i^* to be any maximizing index i.)

When (EC.19) holds, the objective value of (EC.18) can therefore be further refined as

$$\begin{aligned}
&\max\{\mathbf{p}^T\mathbf{b} \mid \sum_{i=1}^{m} p_i Y_i \leq 1, \mathbf{p} \geq \mathbf{0}\}, \\
&= p_{i^*} \cdot b_{i^*} \\
&= \frac{b_{i^*}}{Y_{i^*}} \\
&\leq \frac{\max_{i \in [m]} b_i}{\min_{i \in [m]} Y_i} \\
&\leq \frac{\max_{i \in [m]} b_i}{n \min_{i' \in [m]} q_{i'} - \sqrt{\frac{n}{2} \log \frac{m}{\delta}}} \\
&= \frac{\max_{i \in [m]} b_i}{n} \cdot \frac{1}{\min_{i' \in [m]} q_{i'} - \sqrt{\frac{1}{2n} \log \frac{m}{\delta}}},
\end{aligned}$$

which holds with probability at least $1 - \delta$. This establishes the required bound on the objective value of the infinity norm problem in the theorem statement.

To see that P^{covering} is feasible, observe that $P^{\text{covering}} \equiv \min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is feasible if and only if $\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is feasible. The latter problem is feasible if and only if its dual problem $\max\{\mathbf{p}^T\mathbf{b} \mid \sum_{j=1}^n (\mathbf{p}^T\mathbf{A}_j)_+ \leq 1, \mathbf{p} \geq \mathbf{0}\}$, which is always feasible, is bounded. Our reasoning above establishes that this problem is bounded with probability at least $1-\delta$, which implies that P^{covering} is feasible with probability at least $1-\delta$. This completes the proof. □

We can now complete the proof of Theorem 6.

Proof of Theorem 6: Let β be defined as

$$\beta = \frac{\max_{i\in[m]} b_i}{n} \cdot \frac{1}{\min_{i'\in[m]} q_{i'} - \sqrt{\frac{1}{2n}\log\frac{m}{\delta}}}$$

By Theorem EC.6, we have that P^{covering} is feasible and that $\min\{\|\mathbf{x}\|_\infty \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \leq \beta$ with probability at least $1-\delta$. By Lemma EC.15, observe that by setting C as

$$\begin{aligned} C &= n\beta \\ &= \max_{i\in[m]} b_i \cdot \frac{1}{\min_{i'\in[m]} q_{i'} - \sqrt{\frac{1}{2n}\log\frac{m}{\delta}}}, \end{aligned}$$

we have that $P^{\text{covering}}_{\text{distr}}$ is feasible and

$$\begin{aligned} \Delta v(P^{\text{covering}}_{\text{distr}}) &\leq \sqrt{n}\beta \\ &= \frac{\max_{i\in[m]} b_i}{\sqrt{n}} \cdot \frac{1}{\min_{i'\in[m]} q_{i'} - \sqrt{\frac{1}{2n}\log\frac{m}{\delta}}}, \end{aligned}$$

as desired. □

Appendix D Special Structures and Extensions

In this section, we demonstrate how the results of Sections 3 and B can be applied to LPs with specific problem structures, including LPs with totally unimodular constraints (Section D.1), Markov decision processes (Section D.2), covering problems (Section D.3) and packing problems (Section D.4). In Section D.5, we consider the portfolio optimization problem, which is in general not an LP, but is amenable to the same type of analysis.

D.1 LPs with Totally Unimodular Constraints

Consider a linear program with a totally unimodular constraint matrix, i.e., every square submatrix of $\mathbf{A}$ has determinant 0, 1, or -1. Such LPs appear in various applications, such as minimum cost network flow problems and assignment problems (Bertsekas 1998). In such problems, it is not uncommon to encounter the situation where the number of variables is much larger than the number of constraints. For example, in a minimum cost network flow problem, each constraint

corresponds to a flow-balance constraint at a given node, while each variable corresponds to the flow over an edge; in a graph of n nodes, one will therefore have n constraints and as many as $\binom{n}{2}$ decision variables. We can thus consider solving the problem using the column randomization method. We obtain the following guarantee on the objective value of the column randomization method when applied to linear programs with totally unimodular constraints.

PROPOSITION EC.7. *When* $\mathbf{A}$ *is totally unimodular, then*

$$\gamma = m\|\mathbf{c}\|_\infty$$

is a valid upper bound on $\|\mathbf{p}\|_\infty$ *for every basic solution* $\mathbf{p}$ *of* D.

Proof: Any basic solution $\mathbf{p}$ to the dual problem D can be written as $\mathbf{p}^T = \mathbf{c}_B^T \mathbf{A}_B^{-1}$, where B is a basis. In addition, since $\mathbf{A}$ is totally unimodular, any element of $\mathbf{A}_B^{-1}$ is either 1, -1, or 0. Therefore, the ith component of $\mathbf{p}$ satisfies $p_i = \sum_{j=1}^m [\mathbf{A}_B^{-1}]_{ji}(\mathbf{c}_B)_j \leq m \cdot \|\mathbf{c}\|_\infty$ for all $i \in [m]$. Thus, $\gamma = m\|\mathbf{c}\|_\infty$ is a valid upper bound on $\|\mathbf{p}\|_\infty$. □

Using this result together with the observation that $\|\mathbf{A}\|_{\max} = 1$ for any totally unimodular matrix $\mathbf{A}$, we can invoke Theorem 1 to obtain the following performance guarantee for column randomization when applied to LPs with totally unimodular matrices.

COROLLARY EC.1. *Assume the constraint matrix of* $\mathbf{A}$ *of the complete problem* P *is totally unimodular. Define* C, P_{distr}, P_J *and* $\mathbf{A}_J$ *as in Theorem 1. For any* $\delta \in (0,1)$, *with probability at least* $1-\delta$ *over the set* J, *the following holds: if* P_J *is feasible and* $\operatorname{rank}(\mathbf{A}_J) = m$, *then*

$$\Delta v(P_J) \leq \Delta v(P_{distr}) + \frac{C\,(1+m^2\|\mathbf{c}\|_\infty)}{\sqrt{K}} \left(1 + \sqrt{2\log\frac{2}{\delta}}\right). \tag{EC.20}$$

D.2 Markov Decision Processes

Consider a discounted infinite horizon MDP, with n_s states and n_a actions. The cost function $c(s,a)$ represents the immediate cost of taking action a in state s. The transition probability $P_s(s',a)$ represents the probability of being in state s' after taking action a in state s. Let $\theta \in (0,1)$ be the discount factor. One can solve the MDP by formulating a linear program (Manne 1960):

$$\begin{aligned} \underset{\mathbf{x}_1,\ldots,\mathbf{x}_{n_s} \in \mathbb{R}^{n_a}}{\text{minimize}} \quad & \mathbf{c}_1^T\mathbf{x}_1 + \ldots + \mathbf{c}_s^T\mathbf{x}_s + \ldots + \mathbf{c}_{n_s}^T\mathbf{x}_{n_s} \\ \text{such that} \quad & (\mathbf{E}_1 - \theta\mathbf{P}_1)\mathbf{x}_1 + \ldots + (\mathbf{E}_s - \theta\mathbf{P}_s)\mathbf{x}_j + \ldots + (\mathbf{E}_{n_s} - \theta\mathbf{P}_{n_s})\mathbf{x}_{n_s} = \mathbf{1}, \\ & \mathbf{x}_1, \ldots, \mathbf{x}_s, \ldots, \mathbf{x}_{n_s} \geq \mathbf{0}, \end{aligned}$$

where $\mathbf{E}_j$ is a $n_s \times n_a$ matrix such that the jth row is all ones and every other entry is zero. The vector $\mathbf{c}_s$ is of size n_a such that its ath component is equal to $c(s,a)$. The matrix $\mathbf{P}_s$ is of size

$n_s \times n_a$ such that its (s', a)-th component represents $P_s(s', a)$. Notice that matrix $\mathbf{P}_s$ is a column stochastic matrix, i.e., $\mathbf{1}^T \mathbf{P}_s = \mathbf{1}^T$ and $\mathbf{P}_s \geq \mathbf{0}$ for all $s \in [n_s]$. The decision variable vector $\mathbf{x}_s$ is of size n_a, where the ath entry represents the expected discounted long-run frequency of the system being in state s and action a being taken. If one sorts the decision variables by actions (Ye 2005), then the linear program can be re-written as:

$$\begin{aligned}
\underset{\tilde{\mathbf{x}}^1,\ldots,\tilde{\mathbf{x}}^{n_a} \in \mathbb{R}^{n_s}}{\text{minimize}} \quad & \tilde{\mathbf{c}}_1^T \tilde{\mathbf{x}}_1 + \ldots + \tilde{\mathbf{c}}_a^T \tilde{\mathbf{x}}_a + \ldots + \tilde{\mathbf{c}}_{n_a}^T \tilde{\mathbf{x}}_{n_a} && \text{(EC.21a)} \\
\text{such that} \quad & (\mathbf{I} - \theta \tilde{\mathbf{P}}_1)\tilde{\mathbf{x}}_1 + \ldots + (\mathbf{I} - \theta \tilde{\mathbf{P}}_a)\tilde{\mathbf{x}}_a + \ldots + (\mathbf{I} - \theta \tilde{\mathbf{P}}_{n_a})\tilde{\mathbf{x}}_{n_a} = \mathbf{1}, && \text{(EC.21b)} \\
& \tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_a, \ldots, \tilde{\mathbf{x}}_{n_a} \geq \mathbf{0}, && \text{(EC.21c)}
\end{aligned}$$

where $\tilde{\mathbf{c}}_a = [c(1,a); \ldots; c(s,a); \ldots; c(n_s,a)]$ for $a \in [n_a]$ and $\tilde{\mathbf{P}}_a$ is a $n_s \times n_s$ matrix such that its (s', s)-th element is equal to $P_s(s', a)$. Note that problem (EC.21) is a standard form LP and has more columns than rows. We can therefore apply the column randomization method to solve problem (EC.21). To adapt our performance guarantee from Section 3.2, we establish a bound γ on $\|\mathbf{p}\|_\infty$ for every dual basic solution $\mathbf{p}$ that is specific to problem (EC.21).

PROPOSITION EC.8. *For the infinite horizon discounted MDP problem* (EC.21)*, then*

$$\gamma = \frac{\|\mathbf{c}\|_\infty}{1-\theta}$$

is a valid upper bound on $\|\mathbf{p}\|_\infty$ *for any basic solution* $\mathbf{p}$ *of the dual of problem* (EC.21)*.*

Proof: Any basic solution $\mathbf{p}$ of the dual has the form $\mathbf{p}^T = \mathbf{c}_B^T \mathbf{A}_B^{-1}$, where B is a basis of the linear program (EC.21). Note that $\mathbf{A}_B$ has the form $\mathbf{A}_B = \mathbf{I} - \theta \mathbf{P}$, where $\mathbf{P}$ is an $n_s \times n_s$ matrix such that each of its columns is selected from the columns of $[\tilde{\mathbf{P}}_1, \ldots, \tilde{\mathbf{P}}_{n_a}]$ (see Ye 2005). In addition, a standard property of $\mathbf{A}_B^{-1}$ is that it can be written as the following infinite series:

$$\mathbf{A}_B^{-1} = \mathbf{I} + \theta \mathbf{P} + \theta^2 \mathbf{P}^2 + \cdots = \mathbf{I} + \sum_{n=1}^{\infty} \theta^n \cdot \mathbf{P}^n.$$

Thus, we can bound $\|\mathbf{p}\|_\infty$ as $\|\mathbf{p}^T\|_\infty \leq \|\mathbf{c}_B^T\|_\infty + \sum_{n=1}^{\infty} \theta^n \cdot \|\mathbf{c}_B^T \mathbf{P}^n\|_\infty$. Note that for any $n \in \mathbb{N}$ and vector $\mathbf{v} \in \mathbb{R}^{n_s}$, we have

$$\begin{aligned}
\|\mathbf{v}^T \mathbf{P}^n\|_\infty &= \max_{s \in [n_s]} \left| \sum_{s' \in [n_s]} v_{s'} \mathbf{P}^n_{(s',s)} \right| \\
&\leq \max_{s \in [n_s]} \sum_{s' \in [n_s]} |v_{s'}| \cdot \mathbf{P}^n_{(s',s)} \\
&\leq \|\mathbf{v}\|_\infty \cdot \max_{s \in [n_s]} \sum_{s' \in [n_s]} \mathbf{P}^n_{(s',s)} \\
&= \|\mathbf{v}\|_\infty,
\end{aligned}$$

where $\mathbf{P}^n_{(s',s)}$ is the (s',s)th entry of $\mathbf{P}^n$. Therefore, we obtain that

$$\begin{aligned}\|\mathbf{p}^T\|_\infty &= \|\mathbf{c}_B^T\mathbf{A}_B^{-1}\|_\infty \\ &\leq \|\mathbf{c}_B\|_\infty + \sum_{n=1}^{\infty}\theta^n\cdot\|\mathbf{c}_B^T\mathbf{P}^n\|_\infty \\ &\leq \|\mathbf{c}_B\|_\infty/(1-\theta) \\ &\leq \|\mathbf{c}\|_\infty/(1-\theta).\end{aligned}$$

Since $\mathbf{p}$ was an arbitrary basic solution of the complete dual of problem (EC.21), we can therefore set $\gamma = \|\mathbf{c}\|_\infty/(1-\theta)$. □

With this result in hand, and observing that $\|\mathbf{A}\|_{\max} \leq 1$, we can apply Theorem 1 to obtain the following performance guarantee for column randomization in the case of discounted infinite horizon MDPs.

Corollary EC.2. *Consider solving a discounted infinite horizon MDP with n_s states and n_a actions by the column randomization method. Define C, P_{distr}, P_J and $\mathbf{A}_J$ as in Theorem 1. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds: if P_J is feasible and* $\operatorname{rank}(\mathbf{A}_J) = n_s$*, then*

$$\Delta v(P_J) \leq \Delta v(P_{distr}) + \frac{C}{\sqrt{K}}\cdot\left(1+\frac{n_s\|\mathbf{c}\|_\infty}{1-\theta}\right)\cdot\left(1+\sqrt{2\log\frac{2}{\delta}}\right). \tag{EC.22}$$

D.3 Covering Problems

A covering linear program can be formulated as

$$\begin{aligned} P^{\text{covering}} : \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^T\mathbf{x} && \text{(EC.23a)}\\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \geq \mathbf{b}, && \text{(EC.23b)}\\ & \mathbf{x} \geq \mathbf{0}, && \text{(EC.23c)}\end{aligned}$$

where $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{c}$ are all nonnegative, and we additionally assume that for every $i \in [m]$, there exists a $j \in [n]$ such that $A_{i,j} > 0$. This type of problem arises in numerous applications such as facility location (Owen and Daskin 1998). The column-randomized counterpart of this problem and its dual can be written as

$$\begin{aligned} P_J^{\text{covering}} &: \min\{\mathbf{c}_J^T\tilde{\mathbf{x}} \mid \mathbf{A}_J\tilde{\mathbf{x}} \geq \mathbf{b}, \tilde{\mathbf{x}} \geq \mathbf{0}\},\\ D_J^{\text{covering}} &: \max\{\mathbf{p}^T\mathbf{b} \mid \mathbf{p}^T\mathbf{A}_J \leq \mathbf{c}_J^T, \mathbf{p} \geq \mathbf{0}\}.\end{aligned}$$

Although P^{covering} is not a standard form LP, it is straightforward to extend Proposition EC.3 to this problem, leading to the following result. We omit the proof for brevity.

PROPOSITION EC.9. *Let C be a nonnegative constant and define $P^{covering}_{distr}$ as*

$$P^{covering}_{distr} \equiv \min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}, \mathbf{0} \leq \mathbf{x} \leq C\boldsymbol{\xi}\}.$$

For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, the following holds: if $P^{covering}_J$ is feasible, then

$$\Delta v(P^{covering}_J) \leq \Delta v(P^{covering}_{distr}) + \frac{C}{\sqrt{K}} \cdot (1 + \|\mathbf{p}\|_\infty \cdot m \cdot \|\mathbf{A}\|_{\max}) \cdot \left(1 + \sqrt{2\log\frac{2}{\delta}}\right)$$

for any optimal solution $\mathbf{p}$ of $D^{covering}_J$.

To now use this result, we need to be able to bound $\|\mathbf{p}\|_\infty$ for any solution $\mathbf{p}$ of any dual D^{covering}_J of the column-randomized problem. Let us define the quantity U^{covering} as

$$U^{\text{covering}} = \max_{i,j}\left\{\frac{c_j}{A_{i,j}} \;\middle|\; A_{i,j} > 0\right\}.$$

We then have the following result.

PROPOSITION EC.10. *Let $J \subseteq [n]$, and suppose that $P^{covering}_J$ is feasible. Then for any feasible solution $\mathbf{p}$ of $D^{covering}_J$, $\|\mathbf{p}\|_\infty \leq U^{covering}$.*

Proof: Fix an $i \in [m]$, and consider the LP

$$D^{\text{B}-\text{covering}}_J \;:\; \max\{p_i \mid \mathbf{p}^T\mathbf{A}_J \leq \mathbf{c}_J^T,\; \mathbf{p} \geq \mathbf{0}\}. \tag{EC.24}$$

The optimal objective value of this problem, $v(D^{\text{B}-\text{covering}}_J)$, is an upper bound on p_i for any feasible solution $\mathbf{p}$ of D^{covering}_J (and thus, it is an upper bound on p_i for any optimal solution $\mathbf{p}$ of D^{covering}_J). Consider the dual of this problem:

$$P^{\text{B}-\text{covering}}_J \;:\; \min\{\mathbf{c}_J^T\tilde{\mathbf{x}} \mid \mathbf{A}_J\tilde{\mathbf{x}} \geq \mathbf{e}_i,\; \tilde{\mathbf{x}} \geq \mathbf{0}\}, \tag{EC.25}$$

where $\mathbf{e}_i$ is the ith standard basis vector for $\mathbb{R}^m$. By weak duality, the objective value of any feasible solution of $P^{\text{B}-\text{covering}}_J$ is an upper bound on $v(D^{\text{B}-\text{covering}}_J)$.

We now construct a particular feasible solution. Let j' be any column in J such that $A_{i,j'} > 0$; such a column is guaranteed to exist by our assumption on the matrix $\mathbf{A}$. Define a solution $\tilde{\mathbf{x}}$ as

$$\tilde{x}_j = \begin{cases} 1/A_{i,j'} & \text{if } j = j', \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $\tilde{\mathbf{x}}$ is a feasible solution of $P^{\text{B}-\text{covering}}_J$, and that its objective value is $\mathbf{c}_J^T\tilde{\mathbf{x}} = c_{j'}/A_{i,j'}$. Since this objective value is bounded by U^{covering}, it follows that $U^{\text{covering}} \geq \max\{p_i \mid \mathbf{p}^T\mathbf{A}_J \leq \mathbf{c}_J^T,\; \mathbf{p} \geq \mathbf{0}\}$.

Since our choice of i was arbitrary, it follows that $\|\mathbf{p}\|_\infty \leq U^{\text{covering}}$ for any feasible solution of D^{covering}_J. □

Using this result together with Proposition EC.9 yields the following guarantee.

COROLLARY EC.3. *Let C and $P^{covering}_{distr}$ be defined as in Proposition EC.9. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, the following holds: if $P^{covering}_J$ is feasible, then*

$$\Delta v(P^{covering}_J) \leq \Delta v(P^{covering}_{distr}) + \frac{C}{\sqrt{K}} \cdot (1 + U^{covering} \cdot m \cdot \|\mathbf{A}\|_{\max}) \cdot \left(1 + \sqrt{2 \log \frac{2}{\delta}}\right).$$

D.4 Packing Problems

A packing linear program is defined as

$$\begin{aligned} P^{\text{packing}} : \underset{\mathbf{x}}{\text{maximize}} \quad & \mathbf{c}^T \mathbf{x} && \text{(EC.26a)} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b}, && \text{(EC.26b)} \\ & \mathbf{x} \geq \mathbf{0}, && \text{(EC.26c)} \end{aligned}$$

where we assume that $\mathbf{c} \geq \mathbf{0}$, $\mathbf{b} > \mathbf{0}$, and that $\mathbf{A}$ is such that for every column $j \in [n]$, there exists an $i \in [m]$ such that $A_{i,j} > 0$. Packing problems have numerous applications, such as network revenue management (Talluri and van Ryzin 2006).

The column-randomized counterpart of this problem and its dual can be written as

$$\begin{aligned} P^{\text{packing}}_J &: \max\{\mathbf{c}_J^T \tilde{\mathbf{x}} \mid \mathbf{A}_J \tilde{\mathbf{x}} \leq \mathbf{b}, \tilde{\mathbf{x}} \geq \mathbf{0}\}, \\ D^{\text{packing}}_J &: \min\{\mathbf{p}^T \mathbf{b} \mid \mathbf{p}^T \mathbf{A}_J \geq \mathbf{c}_J^T, \mathbf{p} \geq \mathbf{0}\}. \end{aligned}$$

As with covering problems, the packing problem P^{packing} is not a standard form LP, but we can derive a counterpart of Proposition EC.3 for P^{packing}. Note that in this guarantee, for a problem P' with the same feasible region as P^{packing}, the optimality gap $\Delta v(P')$ is defined as $\Delta v(P') = v(P^{\text{packing}}) - v(P')$, since the complete problem P^{packing} is a maximization problem. As with Proposition EC.9, the proof is straightforward, and thus omitted.

PROPOSITION EC.11. *Let C be a nonnegative constant and define $P^{covering}_{distr}$ as*

$$P^{packing}_{distr} \equiv \max\{\mathbf{c}^T \mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{0} \leq \mathbf{x} \leq C\boldsymbol{\xi}\}.$$

For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, the following holds: if $P^{packing}_J$ is feasible, then

$$\Delta v(P^{packing}_J) \leq \Delta v(P^{packing}_{distr}) + \frac{C}{\sqrt{K}} \cdot (1 + \|\mathbf{p}\|_\infty \cdot m \cdot \|\mathbf{A}\|_{\max}) \cdot \left(1 + \sqrt{2 \log \frac{2}{\delta}}\right)$$

for any optimal solution $\mathbf{p}$ of $D^{packing}_J$.

To obtain a more specific guarantee, define for each i the following quantities:

$$r_i = \max\left\{ \frac{c_j}{A_{i,j}} \;\middle|\; A_{i,j} > 0 \right\},$$
$$j_i^* = \arg\max_j \left\{ \frac{c_j}{A_{i,j}} \;\middle|\; A_{i,j} > 0 \right\}.$$

These two quantities can be understood by interpreting each i as a resource constraint, and b_i as the available amount of resource i. The column j_i^* is the column that has the best rate of objective value garnered per unit of resource i consumed, and the quantity r_i is that corresponding rate. Define now W as

$$W = \sum_{i'=1}^{m} r_{i'} b_{i'},$$

and U^{packing} as the maximum over i of W/b_i, i.e.,

$$U^{\text{packing}} = \max_{i\in[m]} \frac{W}{b_i} = \frac{\sum_{i'=1}^{m} r_{i'} b_{i'}}{\min_{i\in[m]} b_i}.$$

With these definitions, we can establish that U^{packing} is an upper bound on the infinity norm of any dual optimal solution $\mathbf{p}$ of P_J^{packing}.

PROPOSITION EC.12. *Let $J \subseteq [n]$. Then any optimal solution $\mathbf{p}$ of $D_J^{packing}$ satisfies $\|\mathbf{p}\|_\infty \leq U^{packing}$.*

Proof: We first establish a useful property of W: the quantity W is actually an upper bound on $v(P)$. To see this, define the solution $\tilde{\mathbf{x}}^{(i)}$ for each i as

$$\tilde{\mathbf{x}}^{(i)} = \frac{b_i}{A_{i,j_i^*}} \cdot \mathbf{e}_{j_i^*},$$

and define $\tilde{\mathbf{x}} = \sum_{i=1}^{m} \tilde{\mathbf{x}}^{(i)}$. Let $\mathbf{x}$ be any feasible solution of the complete problem P^{packing}. Note that for each $\tilde{\mathbf{x}}^{(i)}$, we have:

$$\begin{aligned}
\mathbf{c}^T \tilde{\mathbf{x}}^{(i)} &= \frac{c_{j_i^*} b_i}{A_{i,j_i^*}} \\
&\geq \frac{c_{j_i^*}}{A_{i,j_i^*}} \left[\sum_{j=1}^{n} A_{i,j} x_j \right] \\
&= \frac{c_{j_i^*}}{A_{i,j_i^*}} \left[\sum_{j:A_{i,j}>0} A_{i,j} x_j \right] \\
&\geq \sum_{j:A_{i,j}>0} A_{i,j} \cdot \frac{c_j}{A_{i,j}} \cdot x_j \\
&= \sum_{j:A_{i,j}>0} c_j x_j.
\end{aligned}$$

where the first inequality follows because $\mathbf{x}$ satisfies $\mathbf{A}\mathbf{x} \leq \mathbf{b}$, and the second follows by the definition of j_i^*. Using this bound, we have

$$\begin{aligned}
\mathbf{c}^T\tilde{\mathbf{x}} &= \sum_{i=1}^{m} \mathbf{c}^T\tilde{\mathbf{x}}^{(i)} \\
&\geq \sum_{i=1}^{m} \left[\sum_{j:A_{i,j}>0} c_j x_j \right] \\
&\geq \sum_{j=1}^{n} c_j x_j \\
&= \mathbf{c}^T\mathbf{x},
\end{aligned}$$

where the second inequality follows by our assumption that for each j, there exists an i such that $A_{i,j} > 0$.

Now, let us fix an $i \in [m]$. We wish to bound $|p_i|$ for an optimal solution $\mathbf{p}$ of D_J^{packing}. We can compute a bound on $|p_i|$ by solving the following LP:

$$D_J^{\text{B}-\text{packing}} \ : \ \max\{p_i \,|\, \mathbf{p}^T\mathbf{b} \leq v(P_J^{\text{packing}}),\ \mathbf{p}^T\mathbf{A}_J \geq \mathbf{c}_J^T,\ \mathbf{p} \geq \mathbf{0}\}.$$

Note that by weak duality, the feasible region of $D_J^{\text{B}-\text{covering}}$ is exactly the set of all optimal solutions to the sampled dual problem, D_J^{packing}. Observe that for any J, $v(P_J^{\text{packing}}) \leq v(P^{\text{packing}}) \leq W$. Thus, a valid upper bound on $v(D_J^{\text{B}-\text{packing}})$ can be obtained by solving the following relaxation of $D_J^{\text{B}-\text{packing}}$:

$$D_J^{\text{B}-\text{packing}-\text{rlx}} \ : \ \max\{p_i \,|\, \mathbf{p}^T\mathbf{b} \leq W,\ \mathbf{p} \geq \mathbf{0}\}.$$

This problem is a valid relaxation, because we have simply removed the constraint $\mathbf{p}^T\mathbf{A}_J \geq \mathbf{c}_J^T$, and we have replaced the value $v(P_J^{\text{packing}})$ with the larger value of W. The optimal objective value of this relaxation is simply W/b_i. Therefore, we obtain that for any dual optimal solution $\mathbf{p}$ of D_J^{packing}, $|p_i| \leq W/b_i$. It follows that $\|\mathbf{p}\|_\infty \leq \max_{i\in[m]}(W/b_i) \equiv U^{\text{packing}}$, for any optimal solution $\mathbf{p}$ of D_J^{packing}. □

By combining this result with Proposition EC.11, we obtain the following specific guarantee for packing LPs.

Corollary EC.4. *Let C and $P_{distr}^{packing}$ be defined as in Proposition EC.11. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, the following holds: if $P_J^{packing}$ is feasible, then*

$$\Delta v(P_J^{packing}) \leq \Delta v(P_{distr}^{packing}) + \frac{C}{\sqrt{K}} \cdot (1 + U^{packing} \cdot m \cdot \|\mathbf{A}\|_{\max}) \cdot \left(1 + \sqrt{2\log\frac{2}{\delta}}\right).$$

With regard to U^{packing} which appears in this guarantee, we note that this constant depends on the constant W. Our choice of W is special only in that it bounds $v(P_J^{\text{packing}})$. For particular packing problems, if one has access to a problem-specific bound W' on $v(P_J^{\text{packing}})$, one could define U^{packing} with W' instead to obtain a more refined bound.

D.5 Portfolio Optimization

In this last section, we deviate slightly from our previous examples by showing how our approach can be applied to problems that are not linear programs. The specific problem that we consider is the portfolio optimization problem, which is defined as

$$
\begin{aligned}
P^{\text{portfolio}}: \quad \underset{\mathbf{x}\in\mathbb{R}^n,\mathbf{r}\in\mathbb{R}^m}{\text{minimize}} \quad & f(r_1,\ldots,r_m) && \text{(EC.27a)}\\
\text{such that} \quad & \sum_{j=1}^{n}\alpha_{ij}x_j = r_i, \quad \forall i\in[m] && \text{(EC.27b)}\\
& \sum_{j=1}^{n} x_j = 1, && \text{(EC.27c)}\\
& \mathbf{x}\geq\mathbf{0}, && \text{(EC.27d)}
\end{aligned}
$$

where both $\mathbf{x}$ and $\mathbf{r}$ are decision variables. Problem (EC.27) can be interpreted as follows: a decision maker seeks an optimal portfolio, which is a distribution over instruments, according to some objectives. The decision variable x_j represents the fraction of allocation committed to instrument j, the constraint parameter α_{ij} represents the return of instrument j in scenario i, and r_i is the total return in ith scenario. The objective function f is a function measuring the risk of the returns $r_1,\ldots,r_m$. Unlike the optimization problems we discussed so far, we assume that f is any Lipschitz continuous function with Lipschitz constant L, and is not necessarily a linear function of $\mathbf{r}$.

Although problem $P^{\text{portfolio}}$ is not in general a linear program, we can still apply the column randomization method to solve the problem. We describe the procedure in Algorithm 9. Notice that, unlike Algorithm 1 which samples columns associated with all variables, here we only sample columns associated with $\mathbf{x}$.

Algorithm 9 The Column Randomization Method - Portfolio Optimization

1: Sample K i.i.d. indices in $[n]$ as $J\equiv\{J_1,\ldots,J_K\}$ according to a randomization scheme ρ.
2: Solve the sampled optimization problem:

$$
P_J^{\text{portfolio}}: \quad \min\left\{ f(\mathbf{r}) \;\middle|\; \sum_{j\in J}\alpha_{ij}\tilde{x}_j = r_i,\ \forall\ i\in[m],\ \sum_{j\in J}\tilde{x}_j = 1,\ \tilde{\mathbf{x}}\geq\mathbf{0} \right\} \qquad \text{(EC.28)}
$$

3: **return** optimal solution $(\tilde{\mathbf{x}}^*,\mathbf{r}^*)$ and optimal objective value $f(\mathbf{r}^*)$

For $P_J^{\text{portfolio}}$ that is produced and solved by Algorithm 9, we have the following performance guarantee.

PROPOSITION EC.13. *Assume vectors* $\boldsymbol{\alpha}_j = (\alpha_{ij})_{i\in[m]}$ *in problem* $P^{portfolio}$ *satisfying* $\|\boldsymbol{\alpha}_j\|_2 \leq H$ *for all* $j \in [n]$. *Let* $C \geq 1$ *be an arbitrary constant and define the optimization problem*

$$P^{portfolio}_{distr}: \quad \min_{\mathbf{x},\mathbf{r}} \left\{ f(\mathbf{r}) \;\middle|\; \sum_{j\in[n]} \boldsymbol{\alpha}_j x_j = \mathbf{r},\ \mathbf{1}^T\mathbf{x} = 1,\ \mathbf{0} \leq \mathbf{x} \leq C\boldsymbol{\xi} \right\}. \tag{EC.29}$$

Denote F, F_{distr}, *and* F_J *as optimal objective values of problems* $P^{portfolio}$, $P^{portfolio}_{distr}$, *and* $P^{portfolio}_{J}$, *respectively. Define* $\Delta F_J \equiv F_J - F$ *and* $\Delta F_{distr} = F_{distr} - F$. *For any* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *the following statement holds:*

$$\Delta F_J \leq \Delta F_{distr} + \frac{CLH}{\sqrt{K}} \left(1 + 3\sqrt{\frac{1}{2}\log\frac{4}{\delta}} \right). \tag{EC.30}$$

While the proof (see below) is similar to that of Proposition EC.3 in the construction of a random solution that is close to the solution of the distributional counterpart problem $P^{\text{portfolio}}_{\text{distr}}$, the main difference is that it relies on Lipschitz continuity, rather than LP duality.

It is worthwhile to point out several aspects about this result and the portfolio optimization problem. First, the portfolio optimization problem (EC.27) is not required to be a convex optimization problem; the objective function f can be non-convex, so long as it is Lipschitz continuous. Second, this result is related to a more specific result from our prior work (Chen and Mišić 2022). In that paper, we consider the problem of estimating the decision forest choice model, which is a probability distribution over a collection of decision trees, and show that by solving an optimization problem over a random sample of trees, one can obtain a gap on the ℓ_1 training error of the model that decays with rate $1/\sqrt{K}$ (Theorem 5 of Chen and Mišić 2022). Proposition EC.13 is a generalization of that result to more general optimization problems outside of choice model estimation, and allows for objective functions more general than those based on ℓ_1 distance.

Proof of Proposition EC.13: Let $(\mathbf{x}^{*0}, \mathbf{r}^{*0})$ be an optimal solution of $P^{\text{portfolio}}_{\text{distr}}$. Consider the solution $(\mathbf{x}', \mathbf{r}')$ defined relative to the sample J:

$$\mathbf{x}' = \frac{1}{K}\sum_{k=1}^{K} \frac{x^{*0}_{j_k}}{\xi_{j_k}} \mathbf{e}_{j_k}, \tag{EC.31}$$

$$\mathbf{r}' = \sum_{j\in[n]} \boldsymbol{\alpha}_j x'_j = \frac{1}{K}\sum_{k=1}^{K} (x^{*0}_{j_k}/\xi_{j_k})\boldsymbol{\alpha}_{j_k}. \tag{EC.32}$$

The significance of $(\mathbf{x}', \mathbf{r}')$ is that we will be able to show that $\mathbf{r}'$ will be close to $\mathbf{r}^{*0}$, and that $f(\mathbf{r}')$ will be close to $f(\mathbf{r}^{*0}) = F_{\text{distr}}$. However, $(\mathbf{x}', \mathbf{r}')$ is not necessarily a feasible solution to problem $P^{\text{portfolio}}$, because $\mathbf{x}'$ will in general not satisfy the unit sum constraint. To turn it into a feasible

solution for problem $P^{\text{portfolio}}$, we consider the solution $(\mathbf{x}'', \mathbf{r}'')$ obtained by normalizing $\mathbf{x}'$ by its sum:

$$\mathbf{x}'' = \frac{\mathbf{x}'}{\mathbf{1}^T \mathbf{x}'}, \tag{EC.33}$$

$$\mathbf{r}'' = \frac{\mathbf{r}'}{\mathbf{1}^T \mathbf{x}'}. \tag{EC.34}$$

Note that $(\mathbf{x}'', \mathbf{r}'')$ is a feasible solution of $P_J^{\text{portfolio}}$.

To understand why we consider $(\mathbf{x}', \mathbf{r}')$ and $(\mathbf{x}'', \mathbf{r}'')$, we show how these two solutions can be used to bound the difference between F_J and F_{distr}. Let $(\mathbf{x}, \mathbf{r})$ be an optimal solution of $P_J^{\text{portfolio}}$. We now bound $F_J - F_{\text{distr}}$ as follows:

$$\begin{aligned} F_J - F_{\text{distr}} &= f(\mathbf{r}) - f(\mathbf{r}^{*0}) \\ &\leq f(\mathbf{r}'') - f(\mathbf{r}^{*0}) \\ &= f(\mathbf{r}'') - f(\mathbf{r}') + f(\mathbf{r}') - f(\mathbf{r}^{*0}) \\ &\leq |f(\mathbf{r}'') - f(\mathbf{r}')| + |f(\mathbf{r}') - f(\mathbf{r}^{*0})| \\ &\leq L\|\mathbf{r}'' - \mathbf{r}'\|_2 + L\|\mathbf{r}' - \mathbf{r}^{*0}\|_2 \end{aligned} \tag{EC.35}$$

where the first step follows by the definitions of $(\mathbf{x}, \mathbf{r})$ and $(\mathbf{x}^{*0}, \mathbf{r}^{*0})$; the second step follows because $(\mathbf{x}'', \mathbf{r}'')$ is a feasible solution of $P_J^{\text{portfolio}}$; the third and fourth step follow by algebra and basic properties of absolute values; and the last step follows by the fact that f is Lipschitz continuous with constant L.

We now proceed to show that $\|\mathbf{r}' - \mathbf{r}^{*0}\|_2$ and $\|\mathbf{r}'' - \mathbf{r}'\|_2$ can be bounded with high probability.

Bounding $\|\mathbf{r}' - \mathbf{r}^{*0}\|_2$: To bound this term, let us define for each $k \in [K]$ the random vector $\mathbf{r}_{j_k}$ as

$$\mathbf{r}_{j_k} = \frac{x_{j_k}^{*0}}{\xi_{j_k}} \boldsymbol{\alpha}_{j_k}.$$

We make three important observations about $\mathbf{r}_{j_1}, \ldots, \mathbf{r}_{j_K}$. First, for each k, the norm of $\mathbf{r}_{j_k}$ is bounded as

$$\|\mathbf{r}_{j_k}\|_2 = \left\| \frac{x_{j_k}^{*0}}{\xi_{j_k}} \boldsymbol{\alpha}_{j_k} \right\|_2 \leq \frac{x_{j_k}^{*0}}{\xi_{j_k}} \cdot \left\|\boldsymbol{\alpha}_{j_k}\right\|_2 \leq \frac{C\xi_{j_k}}{\xi_{j_k}} \cdot H = CH.$$

Second, observe that $\mathbf{r}'$ is just the sample mean of $\mathbf{r}_{j_1}, \ldots, \mathbf{r}_{j_K}$, i.e., $\mathbf{r}' = (1/K)\sum_{k=1}^{K} \mathbf{r}_{j_k}$. Lastly, we observe that the expected value of each $\mathbf{r}_{j_k}$ is

$$\mathbb{E}[\mathbf{r}_{j_k}] = \sum_{j \in [n]: \xi_j > 0} \xi_j \cdot \frac{x_j^{*0}}{\xi_j} \boldsymbol{\alpha}_j$$

$$
\begin{aligned}
&= \sum_{j \in [n]: \xi_j > 0} x_j^{*0} \boldsymbol{\alpha}_j \\
&= \sum_{j \in [n]} x_j^{*0} \boldsymbol{\alpha}_j \\
&= \mathbf{r}^{*0},
\end{aligned}
$$

where the third step uses the fact that $x_j^{*0} = 0$ when $\xi_j = 0$ (by virtue of the constraint $\mathbf{0} \leq \mathbf{x} \leq C\boldsymbol{\xi}$). Therefore, the term $\|\mathbf{r}' - \mathbf{r}^{*0}\|_2$ is just the distance between the sample mean of an i.i.d. collection of random vectors from its expected value, where the ℓ_2 norm of each random vector is bounded. We can therefore invoke Lemma EC.1 to assert that

$$
\|\mathbf{r}' - \mathbf{r}^{*0}\|_2 \leq \frac{CH}{\sqrt{K}} \left(1 + \sqrt{2 \log \frac{2}{\delta}}\right) \tag{EC.36}
$$

with probability at least $1 - \delta/2$.

Bounding $\|\mathbf{r}'' - \mathbf{r}'\|_2$: For this term, observe first that since $\mathbf{r}'' = \mathbf{r}'/(\mathbf{1}^T \mathbf{x}')$, we can re-arrange this to obtain that $\mathbf{r}' = (\mathbf{1}^T \mathbf{x}')\mathbf{r}''$. Let us use s to denote the normalization constant, i.e., $s = \mathbf{1}^T \mathbf{x}'$. We can now bound $\|\mathbf{r}'' - \mathbf{r}'\|_2$ in the following way:

$$
\begin{aligned}
\|\mathbf{r}'' - \mathbf{r}'\|_2 &= \|\mathbf{r}'' - s\mathbf{r}''\|_2 \\
&= |s - 1| \cdot \|\mathbf{r}''\|_2 .
\end{aligned}
$$

We now bound $|s - 1|$. Note that s can be written as

$$
s = \mathbf{1}^T \mathbf{x}' = \frac{1}{K} \sum_{k=1}^{K} \frac{x_{j_k}^{*0}}{\xi_{j_k}} \mathbf{1}^T \mathbf{e}_{j_k} = \frac{1}{K} \sum_{k=1}^{K} \frac{x_{j_k}^{*0}}{\xi_{j_k}}.
$$

Letting $w_k = (x_{j_k}^{*0}/\xi_{j_k})$, we obtain $s = (1/K) \sum_{k=1}^{K} w_k$; in other words, s is the average of K i.i.d. random variables, $w_1, \ldots, w_K$. Note that each w_k has expected value $\mathbb{E}[w_k] = \sum_{j \in [n]: \xi_j > 0} (x_j^{*0}/\xi_j) \cdot \xi_j = \sum_{j \in [n]} x_j^{*0} = 1$; therefore, the term $|s - 1|$ represents how much the sample mean s deviates from its expected value of 1. We also observe that each w_k is contained in the interval $[0, C]$. Therefore, using Hoeffding's inequality, we obtain that

$$
\mathbf{Pr}[|s - 1| > \epsilon] = \mathbf{Pr}[|s - \mathbb{E}[s]| > \epsilon] \leq 2 \cdot \exp\left(-\frac{2K\epsilon^2}{C^2}\right), \tag{EC.37}
$$

for any $\epsilon > 0$; by setting $\epsilon = C\sqrt{\log(4/\delta)/(2K)}$, we obtain that

$$
|s - 1| \leq C \sqrt{\frac{1}{2K} \log \frac{4}{\delta}}, \tag{EC.38}
$$

with probability at least $1 - \delta/2$.

With this bound in hand, let us now bound $\|\mathbf{r}''\|_2$. Observe that

$$\|\mathbf{r}'\|_2 \leq \frac{1}{K} \cdot \sum_{k=1}^{K} \left(\frac{x^{*0}_{j_k}}{\xi_{j_k}} \right) \cdot \|\boldsymbol{\alpha}_{j_k}\|_2 \leq \frac{1}{K} \cdot \sum_{k=1}^{K} \left(\frac{x^{*0}_{j_k}}{\xi_{j_k}} \right) \cdot H = s \cdot H,$$

so it follows that $\|\mathbf{r}''\|_2 = (1/s)\|\mathbf{r}'\|_2 \leq H$. We therefore have that $\|\mathbf{r}'' - \mathbf{r}'\|_2$ satisfies

$$\|\mathbf{r}'' - \mathbf{r}'\|_2 \leq \frac{CH}{\sqrt{K}} \sqrt{\frac{1}{2} \log \frac{4}{\delta}},$$

with probability at least $1 - \delta/2$.

Completing the proof: We now put these two bounds together to complete the bound in (EC.35). Combining inequalities (D.5) and (EC.36) together using the union bound, we have that with probability at least $1 - \delta$,

$$\begin{aligned} F_J - F_{\text{distr}} &\leq L\|\mathbf{r}'' - \mathbf{r}'\|_2 + L\|\mathbf{r}' - \mathbf{r}^{*0}\|_2 \\ &\leq L \cdot \frac{CH}{\sqrt{K}} \sqrt{\frac{1}{2} \log \frac{4}{\delta}} + L \cdot \frac{CH}{\sqrt{K}} \left(1 + \sqrt{2 \log \frac{2}{\delta}} \right) \\ &\leq \frac{CHL}{\sqrt{K}} \left(1 + 3\sqrt{\log \frac{4}{\delta}} \right). \end{aligned}$$

By moving F_{distr} to the right hand side, and subtracting F from both sides, we obtain the desired inequality. □

Appendix E Statistically-Dependent Columns

So far we have assumed that each column in the column-randomized linear program is sampled independently. In this section, we show how this assumption can be relaxed. We state our main performance guarantee in Section E.1. In Section E.2, we consider a specific non-i.i.d. column sampling scheme – *groupwise sampling* – which has natural applications in problems such as Markov decision processes, and apply our guarantee from Section E.1 to this sampling scheme.Finally, in Section E.3, we develop a different type of guarantee for the case when columns are uniformly sampled without replacement.

E.1 Performance Guarantees via Dependency Graph and Forest Complexity

We begin by assuming that the randomization scheme ρ is such that $j_1, \ldots, j_K$ still follow the distribution $\boldsymbol{\xi}$, i.e., $\mathbf{Pr}[j_k = t] = \xi_t$ for $k \in [K]$ and $t \in [n]$, but they are no longer independent. Thus, the indices $j_1, \ldots, j_K$ are no longer an i.i.d. sample from $\boldsymbol{\xi}$, and we require a different set of tools to analyze Algorithm 1 and $\Delta v(P_J)$ in this setting.

To analyze the column randomization method, we will make use of a specific concentration inequality from Liu et al. (2019), which requires specifying the dependence structure of a collection of random variables through a specific type of graph. We thus begin by briefly defining the relevant graph-theoretic concepts.

Given an undirected graph G, we use $V(G)$ to denote the vertices of G, and $E(G)$ to denote the edges of G. Given two vertices $u, v \in V(G)$, the edge between u and v is denoted by $\langle u, v\rangle$. We say that u and v are adjacent if $\langle u, v\rangle \in E(G)$. We say that u and v are non-adjacent if they are not adjacent. For two sets of nodes $U, V \subseteq V(G)$, we say that U and V are non-adjacent if u and v are non-adjacent for every $u \in U$ and $v \in V$. Lastly, a graph G is a *forest* if it does not contain any cycles, and is a *tree* if it does not contain any cycles and consists of a single connected component.

With this definitions, we now define the dependency graph, which is a representation of the dependency structure within a collection of random variables.

Definition EC.1. *(Dependency graph) An undirected graph G is called a* dependency graph *of a set of random variables $X_1, X_2, \ldots, X_K$ if it satisfies the following two properties:*

1. $V(G) = [K]$.
2. *For every $I, J \subseteq [K]$, $I \cap J = \emptyset$ such that I and J are non-adjacent, $\{X_i\}_{i\in I}$ and $\{X_j\}_{j\in J}$ are independent.*

We now introduce the concept of a forest approximation from Liu et al. (2019).

Definition EC.2. *(Forest approximation, Liu et al. (2019)) Given a graph G, a forest F, and a mapping $\phi : V(G) \to V(F)$, we say that (ϕ, F) is a* forest approximation *of G if, for any $u, v \in V(G)$ such that $\langle u, v\rangle \in E(G)$, either $\phi(u) = \phi(v)$ or $\langle \phi(u), \phi(v)\rangle \in E(F)$.*

In words, a forest approximation is a mapping of a general graph G to a smaller forest F that is obtained by merging nodes in G. For a given node $v \in V(F)$, the set $\phi^{-1}(v)$ corresponds to the set of nodes in $V(G)$ that were merged to obtain the node v. Using the notion of a forest approximation, we can now define the forest complexity of a graph G.

Definition EC.3. *(Forest complexity, Liu et al. (2019)) Let $\Phi(G)$ denote the set of all forest approximations of G. Given a forest approximation (ϕ, F), define $\lambda_{(\phi,F)}$ as*

$$\lambda_{(\phi,F)} = \sum_{\langle u,v\rangle \in E(F)} \left(|\phi^{-1}(u)| + |\phi^{-1}(v)|\right)^2 + \sum_{i=1}^{k} \min_{u \in V(T_i)} |\phi^{-1}(u)|^2$$

where $T_1, \ldots, T_k$ is the collection of trees that comprise F. We call $\Lambda(G) = \min_{(\phi,F)\in\Phi(G)} \lambda_{(\phi,F)}$ the forest complexity *of G.*

The forest complexity $\Lambda(G)$ quantifies how much the graph G looks like a forest. Notice that $\Lambda(G) \geq |V(G)|$ for any graph G. In practice, we only need an upper bound on $\Lambda(G)$, rather than its exact value; we refer readers to Liu et al. (2019) for several examples on how $\Lambda(G)$ can be bounded.

Given a dependency graph G for the random indices in the set J, we now bound the optimality gap of the column-randomized linear program.

THEOREM EC.7. *Let C be a nonnegative constant, define P_{distr} as in Theorem 1 and assume the random indices in J follow the dependency graph G with forest complexity $\Lambda(G)$. For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the sample J, the following holds: if P_J is feasible and* $\text{rank}(\mathbf{A}_J) = m$, *then*

$$\Delta v(P_J) \leq \Delta v(P_{distr}) + C \cdot (1 + m\gamma \|\mathbf{A}\|_{\max}) \cdot \left(\sqrt{\frac{K + 2|E(G)|}{K^2}} + \sqrt{\frac{2\Lambda(G)\log(2/\delta)}{K^2}} \right), \quad \text{(EC.39)}$$

where γ and $\|\mathbf{A}\|_{\max}$ are defined as in Theorem 1.

Under the same conditions, with probability at least $1-\delta$ over the sample J, the following holds: if P_J is feasible and $\text{rank}(\mathbf{A}_J) = m$, *then*

$$\Delta v(P_J) \leq \Delta v(P_{distr}) + C \cdot \chi \cdot \left(\sqrt{\frac{K + 2|E(G)|}{K^2}} + \sqrt{\frac{2\Lambda(G)\log(1/\delta)}{K^2}} \right), \quad \text{(EC.40)}$$

where χ is defined as in Theorem 2.

The proof (see below) follows by utilizing the McDiarmid inequality for dependent random variables from Liu et al. (2019). We note that Theorem EC.7 is a generalization of Theorems 1 and 2. If $j_1, j_2, \ldots, j_K$ are independent, then the dependency graph G has no edges, and thus $|E(G)| = 0$ and $\Lambda(G) = K$. Therefore, when each column is generated independently, the upper bounds in Theorem EC.7 are equivalent to the bounds in Theorem 1 and 2.

We close this section by now proving Theorem EC.7. Before we can prove Theorem EC.7, we need to establish two auxiliary results. The first result is the analog of Lemma EC.1 for a collection of possibly dependent random variables, formulated in terms of forest complexity.

LEMMA EC.17. *Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$ be K random vectors with same distribution. Let G be the dependency graph of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$. In addition, assume $\|\mathbf{w}_k\|_2 \leq C$ for $k = 1, \ldots, K$. Let $\bar{\mathbf{w}} = (1/K) \cdot \sum_{k=1}^{K} \mathbf{w}_k$. Then for any $\delta \in (0,1)$, we have, with probability at least $1-\delta$,*

$$\|\bar{\mathbf{w}} - \mathbb{E}\bar{\mathbf{w}}\|_2 \leq C \cdot \left(\sqrt{\frac{K + 2 \cdot |E(G)|}{K^2}} + \sqrt{\frac{2 \cdot \Lambda(G)}{K^2} \cdot \log \frac{1}{\delta}} \right).$$

Proof of Lemma EC.17: Define a space $\mathcal{W} \equiv \{\mathbf{z} \mid \|\mathbf{z}\|_2 \leq C\}$. Consider a scalar function $f : \mathcal{W}^K \to \mathbb{R}$ defined as

$$f(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K) = \left\| \frac{1}{K} (\mathbf{z}_1 + \mathbf{z}_2 + \ldots + \mathbf{z}_K) - \mathbb{E}\bar{\mathbf{w}} \right\|_2$$

For any $k \in [K]$ and any $\mathbf{z}_1, \ldots, \mathbf{z}_k, \ldots, \mathbf{z}_K, \mathbf{z}'_k \in \mathcal{W}$, we have

$$|f(\mathbf{z}_1, \ldots, \mathbf{z}_k, \ldots, \mathbf{z}_K) - f(\mathbf{z}_1, \ldots, \mathbf{z}'_k, \ldots, \mathbf{z}_K)| \leq \frac{\|\mathbf{z}_k - \mathbf{z}'_k\|}{K} \leq \frac{2C}{K}.$$

Therefore, f has the bounded differences property (note that in Liu et al. 2019, this is referred to as the **c**-Lipschitz property; see Definition 2.1 of that paper). By Theorem 3.6 of Liu et al. (2019), for any $\epsilon > 0$, we have

$$\mathbf{Pr}\left[f(\mathbf{w}_1, \ldots, \mathbf{w}_K) - \mathbb{E}f(\mathbf{w}_1, \ldots, \mathbf{w}_K) \geq \epsilon\right] \leq \exp\left(-\frac{K^2\epsilon^2}{2C^2 \cdot \Lambda(G)}\right)$$

On the other hand, define $\mathbf{u}_i = \mathbf{w}_i - \mathbb{E}\mathbf{w}_i$. Then

$$\mathbb{E}\left[\mathbf{u}_i^T \mathbf{u}_j\right] = \begin{cases} \mathbb{E}\left[\mathbf{w}_i^T \mathbf{w}_j\right] - \|\mathbb{E}\mathbf{w}_i\|_2^2 \leq \mathbb{E}\left[\|\mathbf{w}_i\|_2 \|\mathbf{w}_j\|_2\right] \leq C^2, & \text{if } i = j \text{ or } \langle i, j\rangle \in E(G), \\ 0, & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{w}_1, \ldots, \mathbf{w}_K)^2\right] &= \left\| \frac{1}{K}\left(\mathbf{w}_1 + \ldots + \mathbf{w}_K\right) - \mathbb{E}\bar{\mathbf{w}} \right\|_2^2 \\ &= \frac{1}{K^2}\left(\sum_{i,j \in [K]} \mathbb{E}\left[\mathbf{u}_i^T \mathbf{u}_j\right]\right) \\ &= \frac{1}{K^2}\left(\sum_{i \in [K]} \mathbb{E}\left[\mathbf{u}_i^T \mathbf{u}_i\right] + \sum_{\langle i,j\rangle \in E(G)} \mathbb{E}\left[\mathbf{u}_i^T \mathbf{u}_j\right]\right) \\ &\leq C^2 \cdot \frac{K + 2|E(G)|}{K^2}. \end{aligned}$$

As a result,

$$\mathbb{E}f(\mathbf{w}_1, \ldots, \mathbf{w}_K) \leq \sqrt{\mathbb{E}f(\mathbf{w}_1, \ldots, \mathbf{w}_K)^2} \leq C \cdot \sqrt{\frac{K + 2|E(G)|}{K^2}},$$

where the first inequality comes from the concavity of square root function. With all the results above, we have

$$\begin{aligned} \mathbf{P}\left[f(\mathbf{w}_1, \ldots, \mathbf{w}_K) - C \cdot \sqrt{\frac{K + 2|E(G)|}{K^2}} \geq \epsilon\right] &\leq \mathbf{P}\left[f(\mathbf{w}_1, \ldots, \mathbf{w}_K) - \mathbb{E}f(\mathbf{w}_1, \ldots, \mathbf{w}_K) \geq \epsilon\right] \\ &\leq \exp\left(-\frac{K^2\epsilon^2}{2C^2 \cdot \Lambda(G)}\right) \end{aligned}$$

Let $\epsilon = \sqrt{2C^2\Lambda(G)\log(1/\delta)/K^2}$. Then with probability at least $1 - \delta$, we have

$$f(\mathbf{w}_1, \ldots, \mathbf{w}_K) \leq C \cdot \sqrt{\frac{K + 2|E(G)|}{K^2}} + C\sqrt{\frac{2 \cdot \Lambda(G)}{K^2}\log\left(\frac{1}{\delta}\right)}.$$

We thus prove the statement. □

From Lemma EC.17, we can also straightforwardly prove the following result, which is the analog of Lemma EC.2 for possibly dependent random variables.

COROLLARY EC.5. *Let $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$ be K random vectors of size m and with same distribution. Let G be the dependency graph of $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K$. In addition, assume $\|\mathbf{w}_k\|_\infty \leq C$ for $k = 1, \ldots, K$. Let $\bar{\mathbf{w}} = (1/K) \cdot \sum_{k=1}^{K} \mathbf{w}_k$. Then for any $\delta \in (0,1)$, we have, with probability at least $1-\delta$,*

$$\|\bar{\mathbf{w}} - E\bar{\mathbf{w}}\|_1 \leq \sqrt{m} \cdot C \cdot \left(\sqrt{\frac{K + 2 \cdot |E(G)|}{K^2}} + \sqrt{\frac{2 \cdot \Lambda(G)}{K^2} \cdot \log \frac{1}{\delta}} \right).$$

With these two results, we can now proceed with proving Theorem EC.7.

Proof of Theorem EC.7: We define $\mathbf{x}^{*0}$ and construct random vectors $\mathbf{w}_{j_1}, \ldots, \mathbf{w}_{j_K}, \mathbf{b}_{j_1}, \ldots, \mathbf{b}_{j_K}$ as in the proof of Proposition EC.3; we note that this construction is valid even if there exists dependency between the indices j_1, ..., and j_K. We further define $\mathbf{x}'$ as the sample mean of $\mathbf{w}_{j_1}, \ldots, \mathbf{w}_{j_K}$ and $\mathbf{b}'$ as the sample mean of $\mathbf{b}_{j_1}, \ldots, \mathbf{b}_{j_K}$. By Proposition EC.3 and Expression (EC.13), we have

$$\Delta v(P_J) \leq \Delta v(P_{\text{distr}}) + \|\mathbf{x}' - \mathbf{x}^{*0}\|_2 + \|\mathbf{p}_J^*\|_\infty \cdot \|\mathbf{b}' - \mathbf{b}\|_1. \tag{EC.41}$$

By invoking Lemma EC.17, with probability at least $1-\delta$,

$$\|\mathbf{x}' - \mathbf{x}^{*0}\|_2 \leq C \cdot \left(\sqrt{\frac{K + 2 \cdot |E(G)|}{K^2}} + \sqrt{\frac{2 \cdot \Lambda(G)}{K^2} \cdot \log \frac{1}{\delta}} \right). \tag{EC.42}$$

Similarly, by Corollary EC.5, with probability at least $1-\delta$,

$$\|\mathbf{b}' - \mathbf{b}\|_1 \leq \sqrt{m} \cdot C \cdot \|\mathbf{A}\|_{\max} \cdot \left(\sqrt{\frac{K + 2 \cdot |E(G)|}{K^2}} + \sqrt{\frac{2 \cdot \Lambda(G)}{K^2} \cdot \log \frac{1}{\delta}} \right). \tag{EC.43}$$

Combining inequalities (EC.41), (EC.42), and (EC.43) and applying the union bound, we conclude that, with probability at least $1-\delta$, the following holds: if P_J is feasible and $\text{rank}(\mathbf{A}_J) = m$,

$$\Delta v(P_J) \leq \Delta v(P_{\text{distr}}) + C \cdot (1 + m\gamma \|\mathbf{A}\|_{\max}) \cdot \left(\sqrt{\frac{K + 2|E(G)|}{K^2}} + \sqrt{\frac{2\Lambda(G)\log(2/\delta)}{K^2}} \right). \tag{EC.44}$$

Similarly, by Proposition EC.3 and inequality (EC.15), we have

$$\Delta v(P_J) \leq \Delta v(P_{\text{distr}}) + \chi \cdot \|\mathbf{x}' - \mathbf{x}^{*0}\|_2. \tag{EC.45}$$

Combining with inequality (EC.42), we conclude that, with probability $1-\delta$, the following holds: if P_J is feasible and $\text{rank}(\mathbf{A}_J) = m$,

$$\Delta v(P_J) \leq \Delta v(P_{\text{distr}}) + C \cdot \chi \cdot \left(\sqrt{\frac{K + 2|E(G)|}{K^2}} + \sqrt{\frac{2\Lambda(G)\log(1/\delta)}{K^2}} \right), \tag{EC.46}$$

which completes the proof. □

E.2 Groupwise Column Sampling

In many linear programs, we can naturally rearrange and group related columns together. For example, in the LP formulation of an MDP, one can collect columns associated with state s into a set $\mathcal{G}(s)$; the collection of all columns is simply the disjoint union $\bigcup_{s=1}^{n_s} \mathcal{G}(s)$, where n_s is number of states in the MDP and each $\mathcal{G}(s) = \{(s,a) \mid a \in [n_a]\}$. For such a problem, sampling $J = \{j_1, \ldots, j_K\}$ independently from the complete collection of columns, i.e., from $[n_s] \times [n_a]$, may not be attractive. The reason for this is that we may sample the columns in such a way that we do not sample any columns corresponding to a particular state $\tilde{s}$; in such a scenario, the sampled problem P_J will automatically be infeasible.

In the presence of a natural group structure of the columns, rather than sampling columns in total across all n columns, one could consider sampling n_r columns from each group. In the MDP example, this would correspond to sampling n_r columns (which correspond to state-action pairs) for each state s. The resulting column-randomized linear program P_J corresponds to an MDP where there is a random set of n_r actions out of the complete set of n_a actions available in each state s. Most importantly, P_J is guaranteed to be feasible.

It turns out that our results for dependent columns can be used to study column-randomized LPs where columns are sampled by groups. We refer to such a mechanism as a *groupwise randomization scheme* and define it formally below.

Definition EC.4. *(Groupwise Randomization Scheme) Assume the set of indices $[n]$ can be organized into $n_{\mathcal{G}}$ groups, i.e., $[n]$ is the disjoint union of sets $\mathcal{G}_g$ for $g = 1, 2, \ldots, n_{\mathcal{G}}$. Consider a randomization scheme ρ such that (i) it samples indices in n_r rounds of sampling; (ii) in each round, it samples $n_{\mathcal{G}}$ indices as follows: for $i = 1, \ldots, n_{\mathcal{G}}$, it first uniformly at random chooses an index g_i from $[n_{\mathcal{G}}] \setminus \{g_j \mid j \in [i-1]\}$ then samples an index from group $\mathcal{G}_{g_i}$ according to a distribution $\boldsymbol{\xi}^{g_i}$. We refer to such a randomization scheme ρ as a* groupwise randomization scheme.

Note that the randomization scheme ρ samples $K = n_r n_{\mathcal{G}}$ indices in total, and samples n_r columns in each group. By design, each random index j follows the distribution $\boldsymbol{\xi}$, whose probabilities are given by

$$\xi_t \equiv \mathbf{Pr}[j = t] = \frac{1}{n_{\mathcal{G}}} \sum_{g \in [n_{\mathcal{G}}]} \mathbb{I}\{t \in \mathcal{G}_g\} \cdot \xi_t^g = \frac{1}{n_{\mathcal{G}}} \cdot \xi_t^{\mathcal{G}(t)}$$

where $\mathcal{G}(t)$ is the group to which column $t \in [n]$ belongs to.

By using our general result for dependent columns (Theorem EC.7), we obtain a specific guarantee for column-randomized LPs obtained by groupwise randomization schemes.

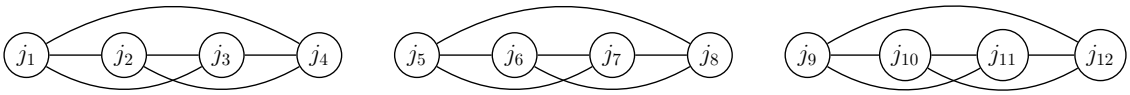

Figure EC.1 Dependency graph of random indices sampled by the groupwise randomization scheme with $n_{\mathcal{G}} = 4$ and $n_r = 3$.

THEOREM EC.8. *Let J be a sample of $K = n_r n_{\mathcal{G}}$ indices sampled according to a groupwise randomization scheme ρ. Let C be a nonnegative constant and define P_{distr} as in Theorem 1. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following holds: if P_J is feasible and $\operatorname{rank}(\mathbf{A}_J) = m$, then*

$$\Delta v(P_J) \leq \Delta v(P_{distr}) + \frac{C\,(1 + m\gamma \|\mathbf{A}\|_{\max})}{\sqrt{n_r}} \left(1 + \sqrt{2 \log \frac{2}{\delta}}\right),$$

where γ and $\|\mathbf{A}\|_{\max}$ are defined as in Theorem 1. Under the same assumption, with probability at least $1-\delta$, the following holds: if P_J is feasible and $\operatorname{rank}(\mathbf{A}_J) = m$, then

$$\Delta v(P_J) \leq \Delta v(P_{distr}) + \frac{C \cdot \chi}{\sqrt{n_r}} \left(1 + \sqrt{2 \log \frac{1}{\delta}}\right),$$

where χ is defined as in Theorem 2.

Proof: The dependency graph G of $K = n_r n_{\mathcal{G}}$ random indices that are sampled by ρ consists of n_r cliques of size $n_{\mathcal{G}}$; Figure EC.1 provides an example of the dependency graph for $n_r = 3$ and $n_{\mathcal{G}} = 4$. Therefore, $|E(G)| = n_r n_{\mathcal{G}}(n_{\mathcal{G}} - 1)/2$ and $\Lambda(G) \leq \lambda(\phi, F) = n_r n_{\mathcal{G}}^2$ for a forest approximation (ϕ, F) that maps each clique in G as a node in F. By upper bounding $\Lambda(G)$ by $n_r n_{\mathcal{G}}^2$ in Theorem EC.7, and using the fact that $K = n_r n_{\mathcal{G}}$, we complete the proof. □

Theorem EC.8 can be interpreted as a guarantee on the optimality gap as a function of the number of columns sampled *per group*: for a groupwise randomization scheme, the gap decreases at a rate of $1/\sqrt{n_r}$, where n_r is the number of columns sampled per group. Compared to Theorem 1 and 2, the rate of convergence in Theorem EC.8 in terms of the *total* number of columns sampled, which is $K = n_r n_{\mathcal{G}}$, is slower; Theorem 1 and 2 both have a rate of $1/\sqrt{K}$, while Theorem EC.8 has a rate of $1/\sqrt{n_r} \equiv \sqrt{n_{\mathcal{G}}/K}$.

E.3 Sampling without replacement

The final extension of our methodology to the non-i.i.d. case that we shall consider is when the columns $\mathbf{A}_{j_1}, \ldots, \mathbf{A}_{j_K}$ are sampled without replacement. For simplicity, we shall restrict our analysis to the case where this sampling is carried out uniformly over the set of columns $[n]$. Stated differently, a random sample of size K drawn uniformly without replacement from $[n]$ is the set of

columns $\{j_1, \ldots, j_K\}$, where $\{j_1, j_2, \ldots, j_n\}$ is a random permutation of the set of columns $[n]$, with all $n!$ permutations having equal probability.

For this sampling method, we begin with an analog of Lemma EC.1. This lemma uses results from the paper of El-Yaniv and Pechyony (2009), which develops a version of McDiarmid's inequality that applies to the sampling without replacement case, and which may be of independent interest to readers.

LEMMA EC.18. *Assume $\mathbf{w}_1, \ldots, \mathbf{w}_n$ are vectors satisfying $\|\mathbf{w}_j\|_2 \leq C$ for $j \in [n]$. Let $\{i_1, \ldots, i_n\}$ be a random permutation of $[n]$ and $\mathbf{z}_j = \mathbf{w}_{i_j}$ for $j \in [n]$. Define $\bar{\mathbf{z}}_K = \sum_{j=1}^{K} \mathbf{z}_j / K$ and $\bar{\mathbf{z}} = \bar{\mathbf{z}}_n = \mathbb{E}[\mathbf{z}_1]$. Then for any $\delta \in (0,1)$, we have, with probability at least $1-\delta$,*

$$\|\bar{\mathbf{z}}_K - \bar{\mathbf{z}}\|_2 \leq \frac{C}{\sqrt{K}} \cdot \left(\sqrt{\frac{n-K}{n-1}} + \sqrt{\frac{2}{H_{n,K}} \log\left(\frac{1}{\delta}\right)} \right), \tag{EC.47}$$

where

$$H_{n,k} \equiv \frac{n - 1/2}{n-K} \cdot \left(1 - \frac{1}{2\max(K, n-K)}\right).$$

Proof: Call $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$. Define the function $f(\mathbf{Z}) = \|\bar{\mathbf{z}}_K - \bar{\mathbf{z}}\|_2$, which is a $(K, n-K)$ *permutation symmetric* function: that is, if we permute the first K or the last $n-K$ vectors of $\mathbf{z}_1, \ldots, \mathbf{z}_n$, the value of $f(\mathbf{Z})$ remains the same.

Given $\mathbf{Z}$, let us use $\mathbf{Z}^{ij}$ to denote the ordered collection that results from swapping the ith and jth vectors in $\mathbf{Z}$. For $i \in \{1, \ldots, K\}$ and $j \in \{K+1, \ldots, n\}$, we then have

$$|f(\mathbf{Z}) - f(\mathbf{Z}^{ij})| \leq \frac{2C}{K},$$

by the triangle inequality. Therefore, by Lemma 2 of El-Yaniv and Pechyony (2009), we have

$$\mathbf{Pr}\left[f(Z) - \mathbb{E}f(Z) \geq \epsilon\right] \leq \exp\left(-\frac{K\epsilon^2}{2C^2} \cdot \frac{(n-1/2)}{(n-K)} \cdot \left(1 - \frac{1}{2\max(K, n-K)}\right)\right).$$

Define $H_{n,k}$ as $\frac{(n-1/2)}{(n-K)} \cdot \left(1 - \frac{1}{2\max(K,n-K)}\right)$. Therefore, the above inequality implies that with probability at least $1-\delta$, we have

$$f(Z) \leq \mathbb{E}f(Z) + \frac{C}{\sqrt{K}} \cdot \sqrt{\frac{2}{H_{n,K}} \log\left(\frac{1}{\delta}\right)}. \tag{EC.48}$$

Now we will bound $\mathbb{E}f(Z)$. We first define $\mathbb{E}[\|\mathbf{z}_1\|^2] = a$ and $\mathbb{E}[\mathbf{z}_1^T \mathbf{z}_2] = b$. Then

$$\mathbb{E}\left[\|\bar{\mathbf{z}}_K - \bar{\mathbf{z}}\|_2^2\right] = \mathbb{E}\left[\bar{\mathbf{z}}_K^T \bar{\mathbf{z}}_K\right] - \bar{\mathbf{z}}^T \bar{\mathbf{z}} = \frac{1}{K} \cdot a + \frac{K-1}{K} \cdot b - \bar{\mathbf{z}}^T \bar{\mathbf{z}}. \tag{EC.49}$$

Notice that when $K = n$, the left-hand side of Equation (EC.49) is zero. This leads to

$$b = \frac{n}{n-1} \cdot \bar{\mathbf{z}}^T \bar{\mathbf{z}} - \frac{a}{n-1}.$$

Plugging this expression for b back into equation (EC.49) and noticing that $a \leq C^2$, we have

$$\mathbb{E}\left[\|\bar{\mathbf{z}}_K - \bar{\mathbf{z}}\|_2^2\right] = \frac{a}{K} \cdot \left(1 - \frac{K-1}{n-1}\right) - \bar{\mathbf{z}}^T\bar{\mathbf{z}} \cdot \left(1 - \frac{n(k-1)}{k(n-1)}\right) \leq \frac{C^2}{K} \cdot \left(1 - \frac{K-1}{n-1}\right)$$

Combining with Jensen's inequality, we have

$$\mathbb{E}f(Z) \leq \sqrt{\mathbb{E}f^2(\mathbf{Z})} = \sqrt{\mathbb{E}\|\bar{\mathbf{z}}_K - \bar{\mathbf{z}}\|_2^2} \leq \frac{C}{\sqrt{K}} \cdot \sqrt{1 - \frac{K-1}{n-1}}. \tag{EC.50}$$

Finally, we plug inequality (EC.50) into inequality (EC.48), which completes the proof. □

Using this lemma, we can now establish an analog of Proposition EC.3. In this proposition, we work with the distributional counterpart problem $P_{\text{distr}}^{\text{wo-rep}} = \min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{0} \leq \mathbf{x} \leq C/n\mathbf{1}\}$, which is the distributional counterpart corresponding to the uniform distribution on $[n]$ (i.e., with $\xi_j = 1/n$ for all $j \in [n]$).

Proposition EC.14. *Let C be a nonnegative constant and define $P_{distr}^{wo\text{-}rep}$ as the linear program $\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{0} \leq \mathbf{x} \leq C/n\}$. Let $Q = \{q_1, \ldots, q_K\} \subset [n]$ be a set of K indices that are sampled uniformly at random from $[n]$ without replacement. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following statement holds: if P_Q is feasible, then*

$$\Delta v(P_Q) \leq \Delta v\left(P_{distr}^{wo\text{-}rep}\right) + \frac{C}{\sqrt{K}} \cdot (1 + \|\mathbf{p}\|_\infty \cdot m \cdot \|\mathbf{A}\|_{\max}) \cdot \left(\sqrt{\frac{n-K}{n-1}} + \sqrt{\frac{2}{H_{n,K}} \log\frac{2}{\delta}}\right),$$

for any optimal dual solution $\mathbf{p}$ of P_Q.

Proof: The proof follows a similar argument for the i.i.d. case (Proposition EC.3). Let $\mathbf{x}^{0*}$ be an optimal solution to $P_{\text{distr}}^{\text{wo-rep}}$. Consider the solution

$$\mathbf{x}' \equiv \frac{1}{K}\sum_{k=1}^{K} n x_{q_k}^{0*} \mathbf{e}_{q_k} \equiv \frac{1}{K} \cdot \sum_{k=1}^{K} \mathbf{z}_k,$$

where $\mathbf{z}_k = n x_{q_k}^{0*} \mathbf{e}_{q_k}$ for $k \in [K]$. We also define $\mathbf{b}' = \mathbf{A}\mathbf{x}'$.

The vectors $\{\mathbf{z}_k\}_{k=1}^K$ have the following properties. First, for all $k \in [K]$, $\mathbb{E}[\mathbf{z}_k] = \mathbf{x}^*$. Second, $\|\mathbf{z}_k\|_2 \leq C$ for all $k \in [K]$ since $0 \leq x_{q_k}^{0*} \leq C/n$. With these properties and recognizing that $\mathbf{x}' = \bar{\mathbf{z}}_K$, we can invoke Lemma EC.18 and assert that, with probability at least $1 - \delta/2$,

$$\|\mathbf{x}' - \mathbf{x}^{0*}\|_2 \leq \frac{C}{\sqrt{K}} \cdot \left(\sqrt{\frac{n-K}{n-1}} + \sqrt{\frac{2}{H_{n,K}} \log\frac{2}{\delta}}\right).$$

With the similar argument in Step 2 of the proof of Proposition EC.3, it can be easily shown that with probability at least $1 - \delta/2$,

$$\|\mathbf{b}' - \mathbf{b}\|_1 \leq \frac{m \cdot C \cdot \|\mathbf{A}\|_{\max}}{\sqrt{K}} \cdot \left(\sqrt{\frac{n-K}{n-1}} + \sqrt{\frac{2}{H_{n,K}} \log\frac{2}{\delta}}\right).$$

With the concentration inequalities in hand, we can bound the objective value of P_Q following the procedure in Step 3 of the proof of Proposition EC.3. □

With this result, the following analog of Theorem 1 can be established for the uniform sampling without replacement case. The proof is identical to Theorem 1 and is omitted for brevity.

THEOREM EC.9. *Let C be a nonnegative constant and define $P^{wo\text{-}rep}_{distr}$ as the linear program $\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{0} \leq \mathbf{x} \leq C/n\}$. Let $Q = \{q_1, \ldots, q_K\} \subset [n]$ be a set of K indices that are sampled uniformly at random from $[n]$ without replacement. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following statement holds: if P_Q is feasible and $\operatorname{rank}(\mathbf{A}_Q) = m$, then*

$$\Delta v(P_Q) \leq \Delta v\left(P^{wo\text{-}rep}_{distr}\right) + \frac{C}{\sqrt{K}} \cdot (1 + \|\mathbf{p}\|_\infty \cdot m \cdot \|\mathbf{A}\|_{\max}) \cdot \left(\sqrt{\frac{n-K}{n-1}} + \sqrt{\frac{2}{H_{n,K}} \log \frac{2}{\delta}}\right),$$

where γ and $\|\mathbf{A}\|_{\max}$ are defined as in Theorem 1.

Alternatively, we can also establish an analog of Theorem 2. To do so, we require an analog of Proposition EC.4, which we formalize below. The proof of this result follows by straightforwardly combining elements of the proof of Proposition EC.4 and Proposition EC.14 above, and is thus omitted.

PROPOSITION EC.15. *Let C be a nonnegative constant and define $P^{wo\text{-}rep}_{distr}$ as the linear program $\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{0} \leq \mathbf{x} \leq C/n\}$. Let $Q = \{q_1, \ldots, q_K\} \subset [n]$ be a set of K indices that are sampled uniformly at random from $[n]$ without replacement. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following statement holds: if P_Q is feasible, then*

$$\Delta v(P_Q) \leq \Delta v\left(P^{wo\text{-}rep}_{distr}\right) + \frac{C}{\sqrt{K}} \cdot \|\mathbf{c}^T - \mathbf{p}^T\mathbf{A}\|_2 \cdot \left(\sqrt{\frac{n-K}{n-1}} + \sqrt{\frac{2}{H_{n,K}} \log \frac{1}{\delta}}\right),$$

for any optimal dual solution $\mathbf{p}$ of P_Q.

Using this proposition, we can then easily obtain the following counterpart of Theorem 2 for the uniform sampling without replacement case.

THEOREM EC.10. *Let C, $P^{wo\text{-}rep}_{distr}$, and Q be as defined in Theorem EC.9. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the following statement holds: if P_Q is feasible and $\operatorname{rank}(\mathbf{A}_Q) = m$, then*

$$\Delta v(P_Q) \leq \Delta v\left(P^{wo\text{-}rep}_{distr}\right) + \frac{C}{\sqrt{K}} \cdot \chi \cdot \left(\sqrt{\frac{n-K}{n-1}} + \sqrt{\frac{2}{H_{n,K}} \log \frac{1}{\delta}}\right),$$

where χ is an upper bound on $\|\bar{\mathbf{c}}\|_2$ for every basic solution of the complete problem P.

We conclude this section by offering a remark on how the bounds we have developed here compare to our earlier bounds for the i.i.d. case. In particular, we focus on Lemma EC.18, which is the main building block of these results. In the i.i.d. case, the counterpart of Lemma EC.18 is Lemma EC.1 (Lemma 4 of Rahimi and Recht (2009)):

$$\|\bar{\mathbf{w}}_K - \bar{\mathbf{w}}\|_2 \leq \frac{C}{\sqrt{K}} \cdot \left(1 + \sqrt{2 \log\left(\frac{1}{\delta}\right)}\right).$$

We numerically compare the bound in Lemma EC.1 ("i.i.d. bound") to that of Lemma EC.18 ("permutation bound") in Figure EC.2 below. We set $\delta = 0.1$, $n = 100$ and vary K. From this figure, we can see that (i) the permutation bound (EC.47) is always tighter than the standard McDiarmid inequality bound, which is under the i.i.d. assumption; and (ii) as K gets closer to n, the improvement becomes larger.

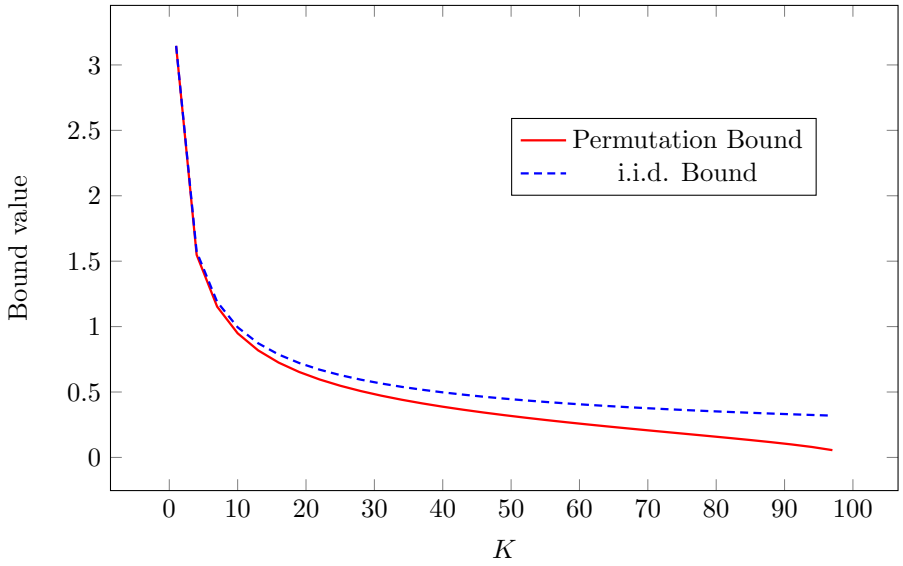

Figure EC.2 Comparison of the bounds of Lemma EC.1 (based on the standard i.i.d. McDiarmid inequality) and Lemma EC.18 (which assumes uniform sampling without replacement), as K varies.

Appendix F Cutting Stock Problem Experiments (continued)

This section continues the numerical experiments with the cutting stock problem in Section 5.

F.1 Experiment #2: comparison of incremental randomization and uniform randomization

In this section, we explore the effect of changing the randomization scheme in the column randomization method. In particular, we compare the incremental randomization scheme ρ_I of the previous section, and the uniform randomization scheme, which we will refer to by ρ_U.

The randomization scheme ρ_U samples from the set $\mathcal{A} = \{\mathbf{a} \in \mathbb{N}_+^m \mid \sum_{i\in[m]} a_i w_i \leq W\}$ uniformly at random. This can be accomplished by rejection sampling. Specifically, we sample uniformly from the set $\bar{\mathcal{A}} = \{\mathbf{a} \in \mathbb{N}_+^m \mid 0 \leq a_i \leq \lfloor W/w_i \rfloor\}$, which can be done by sampling each component a_i uniformly from the set $\{0, 1, \ldots, \lfloor W/w_i \rfloor\}$, and then check if $\sum_{i\in[m]} a_i w_i \leq W$. If this inequality is satisfied, we return $\mathbf{a}$; otherwise, we discard $\mathbf{a}$ and repeat the procedure again with a new candidate column from $\bar{\mathcal{A}}$.

In this experiment, we set $W = 10^5$ again and draw each demand $b_i \sim U(\{1, \ldots, 100\})$ for $i \in [m]$. Due to the poor scaling of rejection sampling, we restrict our focus to the case $m = 5$.

For the widths $w_1, \ldots, w_m$ of the demand types, we consider two different setups:

1. **Setup 1**: we set each width $w_i \sim \text{Uniform}(\{W/1000, W/1000 + 1, W/1000 + 2, \ldots, W/2\})$. In this case, the widths can differ greatly, with two widths w_i and $w_{i'}$ possibly differing by up to a factor of $(W/2)/(W/1000) = 500$.
2. **Setup 2**: we set each width $w_i \sim \text{Uniform}(\{W/10, W/10+1, W/10+2, \ldots, W/4\})$. In this case, the widths are generated to be closer to each other, with two widths w_i and $w_{i'}$ only differing by up to a factor of $(W/4)/(W/10) = 2.5$.

We generate 100 random cutting stock instances in the manner described above for each setup. Then, for a fixed $K \in \{50, 100, 200, 400\}$, we run the column randomization method ten times with each of ρ_U and ρ_I.

Table EC.1 below shows the optimality gap of column randomization with ρ_I and ρ_U under Setups 1 and 2. The reported optimality gap is the average over the 100 cutting stock instances and the ten repetitions of the column randomization method. (Note that for ρ_U in Setup 1, there is no value shown for $K = 50$, as in one replication, the sampled problem was infeasible.) From this table, we can see that when there is high variability in the widths (Setup 1), ρ_I outperforms ρ_U significantly. When there is a lower variability in the widths (Setup 2), ρ_I generally outperforms ρ_U, although the improvement is smaller.

	Setup 1		Setup 2	
K	Δ_{ρ_U} (%)	Δ_{ρ_I} (%)	Δ_{ρ_U} (%)	Δ_{ρ_I} (%)
50	–	1.51	5.80	2.94
100	8.83	0.64	2.57	1.62
200	3.96	0.37	1.13	1.14
400	1.67	0.20	0.28	0.30

Table EC.1 Comparison of ρ_I and ρ_U on the cutting stock problem (Setups 1 and 2).

The edge of ρ_I over ρ_U is not surprising, when one considers the structure of the two randomization schemes. In particular, ρ_I always produces patterns that are maximal, in the sense that no a_i can be further incremented without violating the requirement $\sum_{i=1}^m w_i a_i \leq W$. (Note that such patterns will also be generated by the column generation subproblem (13), as it seeks to maximize $\sum_{i=1}^m p_i a_i$ for some nonnegative dual vector $\mathbf{p}$.) On the other hand, since ρ_U samples uniformly from $\mathcal{A}$, it frequently generates patterns that are inefficient, in that there is space on the large roll for more units of the demand types to be cut from it. Although such patterns can be used to meet the demands, one needs to cut more large rolls according to these patterns (i.e., the corresponding

x_j's need to be larger), resulting in a larger objective value. As a result, ρ_I should yield lower optimality gaps than ρ_U for a fixed K. Nevertheless, this experiment is useful in showing that the choice of randomization scheme is important, and can substantially affect the performance of the column randomization method. We will further underscore this point in our next experiment, where we will also show how the demand vector $\mathbf{b}$ can be used to guide the randomization scheme.

F.2 Experiment #3: comparison of incremental randomization and biased incremental randomization

In this next experiment, we compare incremental randomization with a more sophisticated scheme that we call biased incremental randomization and denote by ρ_{BI}. This new scheme is presented as Algorithm 10. This scheme is the same as the incremental randomization scheme, with the key modification that at each iteration, the index i is sampled with probability proportional to $\sqrt{b_i}$.

The rationale behind this modification is as follows. Suppose that there is significant variability in the demands for different widths, e.g., for a width i, the demand b_i could be very large, but for a different i', the demand $b_{i'}$ could be very small. In such a situation, it may be advantageous to sample patterns where a_i will tend to be large for highly demanded widths, while a_i will tend to be small for less demanded widths, as patterns that are structured in this way are likely to be the most efficient patterns for meeting the demand. (Conversely, it is likely inefficient to use a pattern that yields a few units of the highly demand width and many units of the less demanded width.) The scheme ρ_{BI} is designed exactly for this case, and ensures that the patterns that are produced are such that a_i will be larger when b_i is large.

Algorithm 10 Biased incremental randomization scheme ρ_{BI} for the cutting stock problem.

Column $\mathbf{a}$ is a zero vector of length m and $\zeta \leftarrow W$.
while $\zeta > 0$ **do**
  $I \leftarrow \{i \mid w_i \leq \zeta\}$.
  if $|I| \geq 1$ **then**
    Sample an index i from I with probability $\sqrt{b_i}/\sum_{i \in I} \sqrt{b_i}$.
    Update $a_i \leftarrow a_i + 1$ and $\zeta \leftarrow \zeta - w_i$.
  else
    Break the while loop
return Column $\mathbf{a}$.

In this experiment, we set $W = 10^5$. For each $i \in [m]$, we sample $w_i \sim \text{Uniform}(\{W/10, W/10+1, \ldots, W/4\})$. We then consider two different setups: **Setup 1**, where

each $b_i \sim \text{Uniform}(\{25, \ldots, 100\})$; and **Setup 2**, where $b_i \sim \text{Uniform}(\{50, \ldots, 100\})$. For each setup, we test $m = 50$, $K \in \{100, 150, 200, 250, 300\}$, and $m = 100$, $K \in \{200, 250, 300, 350, 400\}$. For each m and each setup, we generate 100 problem instances, and for each K, we then run column randomization with ρ_I and ρ_{BI} ten times.

Table EC.2 shows the average optimality gap for the two randomization schemes ρ_I and ρ_{BI} under the two different setups. As expected, we can see that in Setup 1, in which the demands exhibit greater variability, the biased scheme ρ_{BI} leads to a lower optimality gap than the ordinary incremental scheme ρ_I. In Setup 2, where there is less variability, ρ_{BI} continues to perform better, although the improvement is smaller.

		Setup 1		Setup 2	
m	K	$\Delta_{\rho_I}(\%)$	$\Delta_{\rho_{BI}}(\%)$	$\Delta_{\rho_I}(\%)$	$\Delta_{\rho_{BI}}(\%)$
50	100	9.41	5.37	6.43	5.90
	150	4.46	2.70	2.63	2.33
	200	2.70	1.59	1.71	1.45
	250	1.66	1.16	1.30	1.21
	300	1.43	0.99	1.00	0.87
100	200	10.46	5.89	6.32	5.34
	250	7.01	3.39	3.78	3.26
	300	4.50	2.30	2.47	2.22
	350	3.18	1.95	2.13	1.82
	400	2.38	1.56	1.61	1.50

Table EC.2 Performance of the randomization schemes ρ_I and ρ_{BI} on the cutting stock problem (Setups 1 and 2).

As we saw in our previous experiment in Section F.1, this experiment illustrates how the choice of randomization scheme can affect the performance of the column randomization method. It also illustrates how the structure of the problem and the nature of the problem data can affect the performance of column randomization and in the same vein, how the problem data can be used in the design of the randomization scheme (in this case specifically, how $\mathbf{b}$ is used in ρ_{BI}).

F.3 Experiment #4: combining column randomization and column generation

In this final experiment, we investigate the potential benefit of combining column randomization and column generation. In particular, we consider a hybrid method, where one first performs column randomization to obtain an initial solution, and then executes column generation starting from that initial solution. The hope in such a method is that column randomization can be used to quickly obtain a good solution with a low optimality gap, and that column generation can then be used to close that gap to zero.

We set up this experiment as follows. We set $W = 10^5$. We vary $m \in \{250, 500, 750, 1000, 1500\}$. For each $i \in [m]$, we draw $w_i \sim \text{Uniform}(\{W/10, \ldots, W/4\})$ and $b_i \sim \text{Uniform}(\{1, \ldots, 100\})$. For simplicity, we set number of sampled columns K in the column randomization as $K = 10m$. For the column randomization method, we use the incremental randomization scheme ρ_I.

Table EC.3 displays the results. The columns labeled "$\Delta_{\rho_{BI}}$ (%)" and "T_{CR} (s)" show the optimality gap of the column randomization solution and the associated computation time. The next column, "$T_{\text{CR-then-CG}}$ (s)", shows the total time required to reach an optimality gap of zero when one executes column generation from the column randomization solution. The last column, "$T_{\text{CG-only}}$ (s)", shows the time required to reach an optimality gap of zero when one applies pure column generation. All values reported are averages over 100 randomly generated instances, and in the case of the columns that involve the column randomization method, are additionally averaged over ten repetitions of the column randomization method.

m	K	$\Delta_{\rho_I}(\%)$	T_{CR} (s)	$T_{\text{CR-then-CG}}$ (s)	$T_{\text{CG-only}}$ (s)
250	2500	1.733	0.403	35.391	52.138
500	5000	1.673	1.774	97.370	130.223
750	7500	1.675	3.496	153.199	241.230
1000	10000	1.630	6.121	268.871	532.083
1500	15000	1.689	14.515	489.697	1107.038

Table EC.3 Performance of the CG method on the cutting stock problem with and without the CR warm start

From this table, we can see that there is a benefit to combining column generation with column randomization. In particular, when $m = 500$, the improvement in the overall time required to reach a 0% gap is small (at most about 30 seconds). For $m = 1000$ and $m = 1500$ the improvement is larger, with the combined method requiring roughly half of the time of the pure column generation method. This experiment illustrates that column randomization can serve as a simple and effective way to obtain an initial solution as an input to column generation, allowing the overall time to be significantly shortened.

F.4 Experiment #5: Exploration of optimal and near-optimal solution for a small instance

In this section, we provide some more insight into why the column randomization method performs well on the cutting stock problem. We consider a small instance with $m = 8$ demand types, with large roll width $W = 200$ and the following widths and demands for the small rolls:

$$\mathbf{w} = (3, 5, 7, 10, 17, 22, 30, 50),$$
$$\mathbf{b} = (1200, 1000, 1000, 400, 500, 400, 600, 200).$$

For this instance, the optimal objective value of problem P^{CS} is 324.5. Our implementation of column generation returns the following solution consisting of 8 patterns that achieves this objective value (note that all x_j values are given to four decimal places):

$$\begin{aligned}
&\mathbf{a}_1 = (0,40,0,0,0,0,0,0), && x_1 = 19.7545\\
&\mathbf{a}_2 = (5,1,0,0,0,0,6,0) && x_2 = 100\\
&\mathbf{a}_3 = (2,0,1,0,11,0,0,0) && x_3 = 45.4545\\
&\mathbf{a}_4 = (0,0,0,20,0,0,0,0) && x_4 = 20\\
&\mathbf{a}_5 = (0,0,0,0,0,0,0,4) && x_5 = 50\\
&\mathbf{a}_6 = (30,0,0,0,0,5,0,0) && x_6 = 18.1371\\
&\mathbf{a}_7 = (0,2,2,0,0,8,0,0) && x_7 = 38.6643\\
&\mathbf{a}_8 = (2,1,27,0,0,0,0,0) && x_8 = 32.4895
\end{aligned}$$

However, this is not the only possible solution. To understand this better, we run the column randomization method 20,000 times, each time with $K = 100$ columns sampled according to ρ_I. For each run of the column randomization method, we solve for the optimal basic feasible solution of the sampled LP and we save the set $S \subseteq \{j_1, \ldots, j_{100}\}$ of patterns for which $x_j > 0$. (Note that while S can contain up to $m = 8$ indices, it could have fewer than 8 indices, because P^{CS} is not a standard form LP.)

Over the 20,000 runs, we obtain 20,000 *unique* sets of columns $S^1, \ldots, S^{20000}$. Of these unique sets, *5946 of them are optimal*. To illustrate, we list below five alternate optimal solutions:

$$\begin{aligned}
\text{Solution 1}: \quad &\mathbf{a}_1 = (3,6,2,2,1,0,2,1) && x_1 = 68.9688\\
&\mathbf{a}_2 = (5,5,2,3,2,1,2,0) && x_2 = 37.4687\\
&\mathbf{a}_3 = (6,1,2,1,3,1,1,1) && x_3 = 59.4375\\
&\mathbf{a}_4 = (3,0,2,3,1,0,1,2) && x_4 = 1.2969\\
&\mathbf{a}_5 = (3,1,5,0,1,2,3,0) && x_5 = 88.3281\\
&\mathbf{a}_6 = (2,4,4,1,2,1,1,1) && x_6 = 26.75\\
&\mathbf{a}_7 = (3,4,1,2,2,0,2,1) && x_7 = 17.3281\\
&\mathbf{a}_8 = (3,3,4,1,0,4,0,1) && x_8 = 24.9219\\
\text{Solution 2}: \quad &\mathbf{a}_1 = (3,1,1,3,1,1,2,1) && x_1 = 36.268\\
&\mathbf{a}_2 = (4,2,3,2,1,0,4,0) && x_2 = 9.0103\\
&\mathbf{a}_3 = (2,6,0,0,0,2,4,0) && x_3 = 81.768
\end{aligned}$$

$\mathbf{a}_4 = (3,3,2,4,0,1,0,2)$ $x_4 = 46.7938$

$\mathbf{a}_5 = (7,3,3,1,3,1,2,0)$ $x_5 = 77.7732$

$\mathbf{a}_6 = (3,1,9,0,3,1,0,1)$ $x_6 = 67.1134$

$\mathbf{a}_7 = (2,3,1,3,4,2,1,0)$ $x_7 = 2.7423$

$\mathbf{a}_8 = (0,2,1,0,3,1,2,1)$ $x_8 = 3.0309$

Solution 3 : $\mathbf{a}_1 = (5,4,0,2,3,2,0,1)$ $x_1 = 21.8947$

$\mathbf{a}_2 = (3,8,4,4,1,3,0,0)$ $x_2 = 22.75$

$\mathbf{a}_3 = (4,2,0,3,4,0,1,1)$ $x_3 = 17.4934$

$\mathbf{a}_4 = (6,2,5,3,1,0,3,0)$ $x_4 = 7.0789$

$\mathbf{a}_5 = (3,2,3,1,0,0,0,3)$ $x_5 = 11.2895$

$\mathbf{a}_6 = (4,4,2,3,0,2,1,1)$ $x_6 = 26.7303$

$\mathbf{a}_7 = (4,3,5,0,2,2,2,0)$ $x_7 = 117.25$

$\mathbf{a}_8 = (3,2,2,1,1,0,3,1)$ $x_8 = 100.0132$

Solution 4 : $\mathbf{a}_1 = (1,1,2,2,2,2,1,1)$ $x_1 = 4.1$

$\mathbf{a}_2 = (2,2,4,1,2,1,3,0)$ $x_2 = 53.1$

$\mathbf{a}_3 = (1,3,3,1,3,0,0,2)$ $x_3 = 10.1$

$\mathbf{a}_4 = (6,4,2,1,2,2,2,0)$ $x_4 = 130.45$

$\mathbf{a}_5 = (2,2,2,1,0,0,2,2)$ $x_5 = 48.95$

$\mathbf{a}_6 = (4,0,6,1,2,1,1,1)$ $x_6 = 16.6$

$\mathbf{a}_7 = (4,3,2,4,1,1,1,1)$ $x_7 = 5.1$

$\mathbf{a}_8 = (2,4,5,2,1,1,1,1)$ $x_8 = 56.1$

Solution 5 : $\mathbf{a}_1 = (7,5,5,0,1,1,1,1)$ $x_1 = 106.8125$

$\mathbf{a}_2 = (2,2,2,1,0,0,2,2)$ $x_2 = 19.375$

$\mathbf{a}_3 = (2,2,4,4,2,1,2,0)$ $x_3 = 44.3438$

$\mathbf{a}_4 = (3,0,0,2,1,2,2,1)$ $x_4 = 3.6875$

$\mathbf{a}_5 = (4,1,2,1,1,1,4,0)$ $x_5 = 54.2812$

$\mathbf{a}_6 = (1,2,1,1,4,1,1,1)$ $x_6 = 50.0625$

$\mathbf{a}_7 = (3,5,0,1,2,1,0,2)$ $x_7 = 0.3437$

$\mathbf{a}_8 = (1,4,2,2,1,3,2,0)$ $x_8 = 45.5937$

There are two important points to note about these solutions. First, notice that all of the columns being used here are very different from the ones used in the column generation solution. In particular, the columns used in the column generation solution are sparser and the magnitudes of the a_i's in those columns are larger. For example, in column $\mathbf{a}_4$ of the CG solution, one cuts 20 units of demand type 4, whereas the most we cut of demand type 4 in any column of the above five alternate solutions is 4 (e.g., column $\mathbf{a}_6$ in solution #5).

Second, the only overlap in the columns used in these five solutions comes from solution #4 and solution #5 (column $\mathbf{a}_5$ of solution #4 is the same as column $\mathbf{a}_2$ of solution #5). Apart from this one column that appears in two of the solutions, every other column only appears once. When we analyze the 5946 optimal solutions that we found, the corresponding columns sets together contain 5123 unique columns (i.e., letting $S^{(i)}$ denote the ith optimal column set, where i ranges from 1 to 5946, we find $|S^{(1)} \cup S^{(2)} \cup \cdots \cup S^{(5946)}| = 5123$). For each column, we calculate its incidence, which is the number of column sets in which the column appears. The maximum incidence of any column is 190, with the average incidence over all of the columns being 9.28 (i.e., on average each column appears in roughly 9 column sets).

Building on the previous statement about the abundance of exactly optimal solutions, there exists an even greater number of *near* optimal solutions. In particular, we can consider the number of solutions that are within $\epsilon = 2.0$ of the optimal objective; note that this absolute gap value translates to a relative gap of $2.0/324.5 = 0.62\%$. We find that that there are 18331 distinct column sets out of the 20,000 that are within $\epsilon = 2.0$ of the optimal objective of 324.5. These distinct column sets span 12294 unique columns, with the maximum incidence of any column being 820 and the average incidence being 11.88.

These last two points, regarding the number of optimal and near-optimal solutions, are important because they directly relate to our analysis of the distributional counterpart in Section 4. In particular, Theorem 3 of Section 4.2 asserts that when there exist many ϵ-optimal BFSs where the incidence of any column is low, then the distributional counterpart gap $\Delta v(P_{\text{distr}})$ will be small. Note that although this result is formulated in terms of BFSs, the same proof technique goes through if one replaces these BFSs with solutions that are supported on a subset of the columns and for which any column appears in at most a certain number of supports. Thus, in the context of the cutting stock problem, it makes sense that column randomization does well, because optimal and nearly-optimal solutions that are diverse in terms of their columns exist in great profusion.

Appendix G Nonparametric Choice Model Estimation Experiments (continued)

This section continues the numerical experiments with the nonparametric choice model estimation in Section 6.

G.1 Experiment #2: Comparison of uniform randomization vs. MNL randomization

In our second experiment, we compare column randomization with the uniform randomization scheme ρ_{Uniform} against column randomization with an alternate randomization scheme that we refer to as *MNL randomization*, and denote by ρ_{MNL}. This randomization scheme involves first fitting an MNL model to the observed choice probabilities and then sampling rankings using the random utility model that underlies MNL. The procedure is formally defined below as Algorithm 11.

Algorithm 11 MNL randomization scheme ρ_{MNL} for the nonparametric choice estimation problem.

Require: Estimated utilities $\hat{u}_1, \ldots, \hat{u}_N$ of each product (via maximum likelihood estimation).

1: Initialize $\alpha_{(i,m)} \leftarrow 0$ for $i \in [N]^+$ and $m \in [M]$.
2: Generate $N+1$ independent random variable $\epsilon_i \leftarrow \text{Gumbel}(0,1)$, for $i \in [N]^+$.
3: Set $v_i \leftarrow \hat{u}_i + \epsilon_i$ for $i \in [N]$, $v_0 \leftarrow 0 + \epsilon_0$.
4: Set σ to be the ranking such that $v_{\sigma(0)} > v_{\sigma(1)} > v_{\sigma(2)} > \ldots > v_{\sigma(N)}$.
5: **for** $m \in [M]$ **do**
6: Set $i^* \leftarrow \arg\min_{i \in S_m \cup \{0\}} \sigma(i)$.
7: Set $\alpha_{(i^*,m)} \leftarrow 1$
8: **return** Column $\boldsymbol{\alpha} = (\alpha_{(i,m)})_{i \in [N]^+, m \in [M]}$.

In this experiment, we again vary N, M and we consider two different setups. In Setup 1, we again sample the utility u_i of each product i as $u_i \sim \text{Uniform}([0,1])$. In Setup 2, we instead sample the utility u_i as $u_i \sim \text{Uniform}([0,20])$.

The rationale for Setup 2 is that when the magnitudes of the utilities $u_1, \ldots, u_N$ are large, then the MNL model begins to behave more and more like a ranking based model. (To see this, suppose that $u_1, \ldots, u_N$ is a collection of distinct non-zero real numbers; observe that for any set $S \subseteq [N]$ and $i \in S$, $\exp(\alpha u_i)/(1 + \sum_{i' \in S} \exp(\alpha u_i')) \to \mathbb{I}\{i = \arg\max_{i' \in S \cup \{0\}} u_{i'}\}$ as $\alpha \to \infty$.) Thus, when the magnitudes of $u_1, \ldots, u_N$ are large, the corresponding distribution over rankings will be concentrated around the ranking that corresponds to $u_1, \ldots, u_N$, i.e., the σ^* such that $u_{\sigma^*(0)} > u_{\sigma^*(1)} > u_{\sigma^*(2)} > \cdots > u_{\sigma^*(N)}$. For this case, we should expect that ρ_{Uniform} will perform poorly, as it is unlikely that we will sample a large number of rankings around σ^*. On the other hand, we should expect ρ_{MNL} to perform better, as it samples rankings from a fitted MNL model whose parameters should be close to the parameters of the true underlying MNL model; thus, ρ_{MNL} should generate rankings that are close to σ^*. Conversely, in Setup 1 (which is identical to our prior setup), we should expect that ρ_{MNL} should improve over ρ_{Uniform}, but the degree of improvement should be smaller. This is because when $u_1, \ldots, u_N \sim \text{Uniform}([0,1])$, the ranking distribution that corresponds to this MNL model will be more diffuse in the space of rankings.

Table EC.4 below shows the performance of column randomization equipped with the two randomization schemes in both Setup 1 and Setup 2. In the table, we use $Z_{\rho_{\text{Uniform}}}$ and $Z_{\rho_{\text{MNL}}}$ to denote the objective value of column randomization equipped with ρ_{Uniform} and ρ_{MNL} respectively. For Setup 1, we can see that in general, ρ_{MNL} does perform better than ρ_{Uniform}; for example, for $N = 8$, $M = 100$, $K = 500$, the objective value of ρ_{MNL} is roughly half of that of ρ_{Uniform}. For Setup 2, the edge of ρ_{MNL} over ρ_{Uniform} is more stark, with ρ_{MNL} resulting in objective values that are 2-3 orders of magnitude smaller than those of ρ_{Uniform}. Note that the two forms of column randomization both have minimal computation time requirements: for ρ_{MNL}, the time to carry out the maximum likelihood estimation, sample the columns, and solve P^{EST} restricted to those columns is no more than 3 seconds across all (N, M, K) combinations. Similarly, the end-to-end computation time for ρ_{Uniform} is also no more than 3 seconds across all (N, M, K) combinations.

			Setup 1		Setup 2	
N	M	K	$Z_{\rho_{\text{Uniform}}}$	$Z_{\rho_{\text{MNL}}}$	$Z_{\rho_{\text{Uniform}}}$	$Z_{\rho_{\text{MNL}}}$
6	50	500	0.06710	0.03665	4.66656	0.06257
		1000	0.00013	0.00244	2.66596	0.02707
8	50	500	0.12338	0.01776	6.89263	0.07756
		1000	0.00000	0.00015	5.24286	0.03840
	100	500	1.04713	0.48256	17.48573	0.14387
		1000	0.20209	0.02196	13.14833	0.09023
		1500	0.00123	0.00103	10.09122	0.05453
10	50	500	0.27030	0.07048	10.14161	0.08363
		1000	0.00003	0.00001	7.85208	0.04030
	100	500	1.51977	0.91140	26.23535	0.10948
		1000	0.35918	0.09530	18.81714	0.07296
		1500	0.03308	0.00034	15.44066	0.05044
		2000	0.00000	0.00027	13.44396	0.04081

Table EC.4 Performance of ρ_{Uniform} and ρ_{MNL} under Setups 1 and 2.

As with our experiments with the cutting stock problem in Sections F.1 and F.2, this experiment illustrates how the structure of the problem data can affect the performance of column randomization: in Setup 2, column randomization equipped with the basic randomization scheme ρ_{Uniform} performs poorly. Simultaneously, this experiment again offers an example of how one can use problem-specific knowledge to design the randomization scheme (in this case, fitting an MNL model, and then sampling from the ranking distribution corresponding to that fitted MNL model). We do acknowledge here that ρ_{MNL} is successful in this experiment because the ground truth model is an MNL model. For other ground truth models (e.g., the nested logit model or the latent-class MNL model), we should no longer expect ρ_{MNL} to do as well. However, for a different type of ground

truth model, one can take the same strategy as in Algorithm 11 where one estimates a different random utility maximization model. (For example, one could fit a latent-class MNL model using expectation-maximization, and then sample rankings from the resulting model.)

Lastly, we also note here that this experiment is congruent with our theoretical results on the distributional counterpart gap under generative model 1 (namely Theorem 4). Recall that in that generative model, the right-hand side is generated as a scaled random convex combination of the set of columns, where the vector of convex combination weights $\boldsymbol{\theta}$ is drawn uniformly from the $(n-1)$-dimensional unit simplex. Although problem P^{EST} is not a standard form LP, there is a similarity here as the right hand side vector $\mathbf{v}$ can also be thought of as being generated by a certain random combination of the columns in $\mathbf{A} = [\boldsymbol{\alpha}_1 \cdots \boldsymbol{\alpha}_{(N+1)!}]$, with the scale factor η of generative model 1 being equal to 1 (see also the discussion in Section 4.3 around the moment estimation problem EC.17). Thus, $\boldsymbol{\theta}$ can be thought of as the true underlying distribution over rankings. Theorem 4 tells us that when $\boldsymbol{\theta}$ is drawn uniformly, and the randomization scheme is such that one samples columns uniformly over $[n]$, then most of the time the gap should be $O(\log n/\sqrt{n})$. The underlying distribution over rankings that one obtains under Setup 1 is closer to looking like a $\boldsymbol{\theta}$ drawn under generative model 1 than the same distribution obtained under Setup 2. This, in turn, explains why the performance of ρ_{Uniform} deteriorates so much from Setup 1 to Setup 2.

G.2 Experiment #3: combining column randomization and column generation

In this final experiment, analogously to the experiment in Section F.3 for the cutting stock problem, we examine the value of using column randomization as a way of warm-starting column generation. We test the same values of (N, M, K) as in Section 6. For each N and M, we generate 100 problem instances in the same manner as in Section 6. Then, for each K, we run the column randomization method equipped with ρ_{Uniform} ten times, and we then use each solution as the initial solution for column generation, which we run until we reach an optimality gap of zero.

Table EC.5 shows the results of this experiment. The columns labeled $Z_{\rho_{\text{Uniform}}}$ and T_{CR} indicate the objective value of the column randomization solution and the time required by column randomization, respectively. The next column, $T_{\text{CR-then-CG}}$, indicates the overall time required for the combined method (column randomization followed by column generation) to reach a zero optimality gap. The last column, $T_{\text{CG-only}}$, shows the time required for ordinary column generation (i.e., without any warm starting) to reach an optimality gap of zero. From this table, we again see that using column randomization to warm start column generation can dramatically reduce the time required to reach an optimality gap of zero.

N	M	K	$Z_{\rho_{\text{Uniform}}}$	T_{CR} (s)	$T_{\text{CR-then-CG}}$ (s)	$T_{\text{CG-only}}$ (s)
6	50	500	0.02990	0.06	3.94	27.55
		1000	0.00093	0.08	0.52	27.55
8	50	500	0.17719	0.17	21.77	107.27
		1000	0.00000	0.15	0.33	107.27
	100	500	0.96841	0.32	397.90	705.70
		1000	0.19307	0.44	222.76	705.70
		1500	0.00000	0.92	1.63	705.70
10	50	500	0.32513	0.26	86.51	284.65
		1000	0.00008	0.24	2.09	284.65
	100	500	1.48889	0.36	1441.17	2311.02
		1000	0.30858	0.53	737.89	2311.02
		1500	0.01132	1.06	133.56	2311.02
		2000	0.00000	2.60	3.87	2311.02
	150	500	2.92142	0.82	6849.05	9831.54
		1000	1.16588	0.98	5358.42	9831.54
		1500	0.47822	1.44	3897.58	9831.54
		2000	0.15500	2.25	1987.44	9831.54
		2500	0.00000	5.82	10.41	9831.54

Table EC.5 Performance of the CG method on the nonparametric choice model estimation problem with and without the CR-based warm start.

Appendix H Comparisons to Other Approaches

We complement Section 2 and make an additional comparison between our work and the two works in the literature.

H.1 Comparison with Agrawal et al. (2014)

The starting point of Agrawal et al. (2014) is an *online* linear program, which is an online version of the following problem:

$$P_{\text{OLP}}: \quad \underset{\mathbf{x}}{\text{maximize}} \quad \sum_{j=1}^{n} \pi_j x_j \tag{EC.51a}$$

$$\text{subject to} \quad \sum_{j=1}^{n} a_{ij} x_j \le b_i, \quad \forall i \in [m], \tag{EC.51b}$$

$$0 \le x_j \le 1, \quad \forall j \in [n]. \tag{EC.51c}$$

At the beginning, the decision maker has no information about the ground truth model, except knowing the total number of columns n. Time progresses in discrete periods, and at each period t, nature randomly reveals a column $\mathbf{a}_t$ and a coefficient π_t to the decision maker. The decision maker then makes a decision x_t based on the history $(\pi_1, \mathbf{a}_1, x_1, \pi_2, \mathbf{a}_2, x_2, \ldots, \pi_t, \mathbf{a}_t)$. Note that this is an *irrevocable* decision: the decision maker cannot change the decisions made in earlier periods later.

The decision maker's goal is to maximize the cumulative reward $\sum_{t=1}^{n} \pi_t x_t$. A critical assumption in Agrawal et al. (2014) is that the arrival order of columns $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ is uniformly distributed over all the permutations (Assumption 1.1 of that paper) and n is known (Assumption 1.2 of that paper).

To solve this online linear program, Agrawal et al. (2014) propose an algorithm called the one-time-learning algorithm (OLA). This algorithm observes the first $s = \lceil \epsilon n \rceil$ periods of time, where $\epsilon \in (0, 1)$, and using the s columns observed, it defines a policy for making decisions for the remaining $n - s$ periods/columns. In particular, one considers following primal and dual problem pair defined on the first s columns:

$$
\begin{aligned}
P_{\mathrm{OLA}}: \quad \underset{\mathbf{x}}{\text{maximize}} \quad & \sum_{t=1}^{s} \pi_t x_t && \text{(EC.52a)}\\
\text{subject to} \quad & \sum_{t=1}^{s} a_{it} x_t \leq (1-\epsilon)\frac{s}{n} b_i, \quad \forall i \in [m], && \text{(EC.52b)}\\
& 0 \leq x_t \leq 1, \quad \forall t \in [s]. && \text{(EC.52c)}
\end{aligned}
$$

$$
\begin{aligned}
D_{\mathrm{OLA}}: \quad \underset{\mathbf{p},\mathbf{y}}{\text{minimize}} \quad & \sum_{i=1}^{m} (1-\epsilon)\frac{s}{n} b_i p_i + \sum_{t=1}^{s} y_t && \text{(EC.53a)}\\
\text{subject to} \quad & \sum_{i=1}^{m} a_{it} p_i + y_t \geq \pi_t, \quad \forall t \in [s], && \text{(EC.53b)}\\
& p_i \geq 0, \quad \forall i \in [m], && \text{(EC.53c)}\\
& y_t \geq 0, \quad \forall t \in [s]. && \text{(EC.53d)}
\end{aligned}
$$

Given a dual vector $\mathbf{p}$ for the dual problem D_{OLA}, define a policy $\mathbf{x}(\cdot)$ as

$$
x_t(\mathbf{p}) = \begin{cases} 0, & \text{if } \pi_t \leq \mathbf{p}^T \mathbf{a}_t, \\ 1, & \text{if } \pi_t > \mathbf{p}^T \mathbf{a}_t. \end{cases} \qquad \text{(EC.54)}
$$

The OLA algorithm then operates as follows:

1) Initialize $x_t = 0$ for all $t \leq s$. Let $\hat{\mathbf{p}}$ be the optimal solution to the dual problem D_{OLA}.
2) For $t = s+1, s+2, \ldots, n$, if $a_{it} x_t(\hat{\mathbf{p}}) \leq b_i - \sum_{j=1}^{t-1} a_{ij} x_j$ for all $i \in [m]$, set $x_t = x_t(\hat{\mathbf{p}})$; otherwise, set $x_t = 0$. Output x_t.

Having provided this overview of the problem setup and the method of Agrawal et al. (2014), a number of critical differences become apparent.

First, the problem setup in Agrawal et al. (2014) is different from that of our paper. Agrawal et al. (2014) considers an online problem: their problem is a sequential decision making problem such that at each time t, the decision maker receives information (a column) $\mathbf{a}_t$, and then makes a

decision x_t. Decisions made in the past cannot be changed. The decision maker cannot know more about the problem until nature reveals more information. In contrast, our problem is not an online problem: the problem is a static problem.

Second, the source of randomness is different. In the model of Agrawal et al. (2014), at each time step t, nature reveals a column $\mathbf{a}_t$ uniformly at random from the remaining columns; thus, the randomness is an inherent part of the problem. The OLA method is deterministic, that is to say, it does not introduce (additional) randomness to solve the problem. In other words, Agrawal et al. (2014) uses a deterministic method to solve a stochastic problem, which is the online linear program. In contrast, in our paper, the ground truth model/problem is a large-scale linear program. This problem is a deterministic, one-shot problem – there is no randomness in how information is revealed to the decision maker, and the decision maker does not need to set decision variables sequentially/in real time – but it is *very large*. Therefore, we introduce randomness in the solution method, i.e., we propose a randomized algorithm to solve this large-scale deterministic problem.

Third, notwithstanding the difference in problem setups, Agrawal et al. (2014) comment on the possibility of using OLA as an offline method to solve large-scale linear programs (see Section 5.3 of that paper). In particular, one first creates a random order of the n columns, samples $s = \lceil \epsilon n \rceil$ columns, solves D_{OLA}, and sets the variables according to the procedure given above. However, this approach is difficult to apply in the setting that we study. First, our linear program is a standard form LP of the form $\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$; in particular, the constraints are not inequalities, and the variables do not have a priori upper bounds, so it is not straightforward to adapt the variable-setting procedure of OLA to this more general problem. Second, even if one can overcome this difficulty, OLA fundamentally requires one to iterate through *all* n columns. This is impossible when n is astronomically large. As an example, in the cutting stock problem that we study in Section 5 of our paper, OLA would involve sampling a small set of s patterns, solving a problem to obtain dual variables, and iterating through *every remaining pattern* to set x_j for those patterns according to the dual variables. Although OLA could be useful for solving offline LPs where n is moderately large – i.e., the full LP is tedious to solve, but solvable – we do not believe that it is computationally feasible for the case where n is so large that the full LP itself cannot be formed and solved directly. This latter setting is precisely the setting that our method is intended for.

H.2 Comparison with Vu et al. (2018)

The approach of Vu et al. (2018) involves reducing the number of constraints in a linear program. In particular, instead of solving the problem $\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$, one forms a random k-by-m matrix $\mathbf{T}$ and left-multiplies both sides of the constraint $\mathbf{A}\mathbf{x} = \mathbf{b}$ by this matrix to obtain the following simplified problem:

$$\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{T}\mathbf{A}\mathbf{x} = \mathbf{T}\mathbf{b}, \mathbf{x} \geq \mathbf{0}\}. \tag{EC.55}$$

This problem has fewer constraints (k constraints, compared to m constraints in the original problem).

Having given an overview of the random projection method, it is clear that there are a number of important differences. First, while our method involves reducing the number of columns by drawing a random sample of columns, the random projection method of Vu et al. (2018) involves reducing the number of constraints by taking a linear combination of the constraints. This is important because in our problem setting, the number of columns n is assumed to be much larger than the number of rows m; thus, random projection does not make the problem simpler to solve.

Second, by replacing the constraint $\mathbf{TAx} = \mathbf{Tb}$, the solution $\mathbf{x}$ may not be feasible for the original equality constraint $\mathbf{Ax} = \mathbf{b}$. In fact, a result of Vu et al. (2018), Proposition 3, asserts that a solution to problem (EC.55) is infeasible for the original problem with probability 1. This contrasts with our setup, where if the sampled problem P_J is feasible, the resulting solution is feasible for the complete problem P; and additionally, one can augment the sampled set of columns J with a set of columns J_F to guarantee feasibility of the sampled set (see Algorithm 7 in Section 3.3). Although Vu et al. (2018) provide a procedure (Algorithm 1) for retrieving an optimal basic feasible solution under certain conditions with high probability, the probability bound scales like $1 - O(n)$, which for our setting where n is extremely large would yield a low probability. Indeed, the authors of Vu et al. (2018) acknowledge finding "very high errors" in applying this retrieval procedure in their numerical experiments (see the discussion in Section 7.2 of Vu et al. 2018), and for this reason consider a heuristic modification of their retrieval algorithm. Thus, guaranteeing a feasible solution to the original problem when applying the random projection method is not a triviality.

Third, a tacit assumption in Vu et al. (2018) is that one can form the matrix $\mathbf{A}$ explicitly, and can carry out the multiplication $\mathbf{TA}$ exactly. This will in general be impossible for the regime that we are interested, where n can be astronomically large. For example, in the cutting stock example we consider, one would need to form the matrix $\mathbf{A}$ containing columns for *all* possible patterns, and then compute $\mathbf{TA}$; without even getting to the question of how one solves the sketched problem (EC.55), forming $\mathbf{A}$ and then $\mathbf{TA}$ is clearly computationally infeasible. For this reason, the numerical examples that are considered in Vu et al. (2018) are of a much smaller scale than the ones we consider: n is at most 2400, and the largest computation time reported for solving the original LP $\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}$ is no more than two minutes.

Lastly, we comment that while the random projection method as originally described in Vu et al. (2018) reduces the number of rows in the primal LP, it is tempting to consider an alternate application of this method where one reduces the number of rows in the dual LP. The dual LP is

$$D : \max\{\mathbf{p}^T\mathbf{b} \mid \mathbf{p}^T\mathbf{A} \leq \mathbf{c}^T\}$$

We can transform this into an equality constrained problem by introducing the slack vector $\mathbf{s} \in \mathbb{R}^n$:

$$D' : \max\{\mathbf{p}^T\mathbf{b} \mid \mathbf{p}^T\mathbf{A} + \mathbf{s}^T = \mathbf{c}^T, \mathbf{s} \geq \mathbf{0}\}.$$

Now, we can right-multiply each side of the inequality constraint by a n-by-k matrix $\mathbf{T}$, where $k \ll n$, resulting in the projected dual problem:

$$D'_{\text{RP}} : \max\{\mathbf{p}^T\mathbf{b} \mid \mathbf{p}^T\mathbf{A}\mathbf{T} + \mathbf{s}^T\mathbf{T} = \mathbf{c}^T\mathbf{T}\}.$$

The dual of this projected dual problem is

$$P_{\text{RP}} : \min\{\mathbf{c}^T\mathbf{T}\tilde{\mathbf{x}} \mid \mathbf{A}\mathbf{T}\tilde{\mathbf{x}} = \mathbf{b}, \mathbf{T}\tilde{\mathbf{x}} \geq \mathbf{0}\},$$

where $\tilde{\mathbf{x}}$ is now a k-dimensional vector of decision variables (versus an n-dimensional vector in the original problem). Although this approach seems promising, again one runs into computation issues. In terms of computation, the matrix $\mathbf{T}$ is enormous as it has n rows, and one needs to carry out the matrix multiplication $\mathbf{AT}$, which for large-scale applications like cutting stock will be impossible. Additionally, although P_{RP} achieves a reduction in the number of decision variables from n to k, there are still $O(n)$ constraints due to the constraint $\mathbf{T}\tilde{\mathbf{x}} \geq \mathbf{0}$. Lastly, a serious limitation of problem P_{RP} is that it may be infeasible. (Comparing problem P_{RP} and the original problem P, P_{RP} is the same as P with the constraint that $\mathbf{x}$ lies in the lower dimensional subspace $\{\mathbf{T}\tilde{\mathbf{x}} \mid \tilde{\mathbf{x}} \in \mathbb{R}^k\}$. With this additional constraint, it is not guaranteed that we can satisfy the equality constraint $\mathbf{A}\mathbf{x} = \mathbf{b}$ and the nonnegativity constraint $\mathbf{x} \geq \mathbf{0}$.)

We can therefore see that even applying random projection in an alternate fashion is problematic for the large-scale LP setting that we study.