APPENDIX

A SAMPLING MULTIPLE KNOCKOFFS

A.1 Gaussian Multi-knockoffs

We generalize the knockoff generation procedure to have $\kappa \geq 2$ multi-knockoffs, starting with the Gaussian case. We see that a sufficient condition for $(\boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa) \in \mathbb{R}^{d\kappa}$ to be a multi-knockoff vector - besides all vectors $\boldsymbol{X}^j$ having the same mean μ- is that the joint vector $(\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa) \in \mathbb{R}^{d(\kappa+1)}$ has a covariance matrix of the form:

$$\Sigma_\kappa = \underbrace{\begin{pmatrix} \Sigma & \Sigma - D & \dots & \Sigma - D \\ \Sigma - D & \Sigma & \dots & \Sigma - D \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma - D & \Sigma - D & \dots & \Sigma \end{pmatrix}}_{\kappa+1 \text{ blocks}}$$

We can easily generalize previous diagonal matrix constructions to the multi-knockoff setting. The mathematical formulation of the heuristic behind SDP and equicorrelated knockoffs -as an objective function in the convex optimization problem- does not change when sampling multi-knockoffs, as the correlation between an original feature and any of its multi-knockoffs is the same as a consequence of exchangeability. However, the positive semi-definite constraint that defines the feasible set changes with κ. For the entropy knockoffs, the objective function depends also on κ.

Because all three methods solve a similar convex optimization problem, there is no significant difference in runtime.

Proposition A.1. *We generalize the diagonal construction methods SDP, equicorrelated and entropy when sampling $\kappa \geq 2$ multi-knockoffs from a multivariate Gaussian, by the following convex optimization problems. We recover the formulations for the single knockoff setting by replacing $\kappa = 1$.*

- ***SDP Multi-knockoffs*** *For a covariance matrix Σ whose diagonal entries are equal to one, the diagonal matrix $D(s) = diag(s_1, \dots, s_d)$ for constructing SDP knockoffs is given by the following convex optimization problem:*

$$\textit{minimize} \quad \sum_{i=1}^{d} |1 - s_i|$$
$$\textit{subject to} \begin{cases} \frac{\kappa+1}{\kappa}\Sigma - D(s) \succ 0 \\ s_i \geq 0 \qquad \forall i \in \{1, \dots, d\} \end{cases}$$

- ***Equicorrelated Multi-knockoffs*** *For a covariance matrix Σ whose diagonal entries are equal to one, the diagonal matrix $D(s) = sI_d$ for constructing equicorrelated knockoffs is given by the following convex optimization problem:*

$$\textit{maximize} \quad s \qquad \textit{subject to} \begin{cases} \frac{\kappa+1}{\kappa}\Sigma - sI_d \succ 0 \\ s \geq 0 \end{cases}$$

  The solution of this optimization problem has a closed form expression: $s^* = \frac{\kappa+1}{\kappa}\lambda_{min}(\Sigma)$, where $\lambda_{min}(\Sigma)$ is the smallest (positive) eigenvalue of Σ.

- ***Entropy Multi-knockoffs*** *The diagonal matrix $D(s) = diag(s_1, \dots, s_d)$ for constructing entropy knockoffs is given by the following convex optimization problem (as $s \mapsto -\log\det(2\Sigma - D(s))$ is convex):*

$$\arg\min_s \; -\log\det(\frac{\kappa+1}{\kappa}\Sigma - D(s)) - \kappa\sum_{i=1}^{d}\log(s_i)$$
$$\textit{subject to} \begin{cases} \frac{\kappa+1}{\kappa}\Sigma - D(s) \succ 0 \\ s_i \geq 0 \qquad \forall i \in \{1, \dots, d\} \end{cases}$$

The entropy knockoff construction method avoids solutions where diagonal terms are extremely close to 0*, and we provide the following lower bound on the diagonal terms of D:*

$$(\lambda_{min}(s))^{\frac{1}{\kappa-2}} \leq \min_{j \in \{1,\dots,d\}} s_j$$

where $\lambda_{min}(s)$ is the smallest (positive) eigenvalue of $\frac{\kappa+1}{\kappa}\Sigma - D(s)$.

For the SDP method and the equicorrelated method, increasing the number of multi-knockoffs constrains the feasible set of the convex optimization problem. However, diagonal terms can always be as close to 0 as they want, and we empirically observe a slight decrease in power as we increase κ indicating that the added constraints limit the choice of "good" values for the diagonal terms.

Proof. The heuristic behind the different construction methods looks for different optimal solutions to convex optimization problems. Depending on the multi-knockoff parameter κ, we need to adapt two parts of the convex optimization formulations: the objective function and the feasible set. Objective functions in the SDP and equicorrelated constructions remain unchanged as they do not depend on the number of multi-knockoffs.

Adapting the Feasible Set We first look at how the constraints defining the feasible set change as we go from simple knockoffs to multi-knockoffs. All three methods (SDP, equicorrelated, entropy) define the feasible set for $s = (s_1, \dots, s_d) \in \mathbb{R}_+^d$ by constraining Σ_κ to be positive definite. We show that this constraint is equivalent to $\frac{1+\kappa}{\kappa}\Sigma - D \succ 0$, which we prove by induction. Suppose that at step $\kappa \geq 1$, for any positive definite matrix S, for D positive definite diagonal matrix,

$$\underbrace{\begin{pmatrix} S & S-D & \dots & S-D \\ S-D & S & \dots & S-D \\ \vdots & \vdots & \ddots & \vdots \\ S-D & S-D & \dots & S \end{pmatrix}}_{\kappa+1 \text{ blocks}} \succ 0$$

$$\Leftrightarrow \qquad \frac{1+\kappa}{\kappa} S - D \succ 0$$

where we write $A \succ 0$ for A symmetric positive definite. We repeatedly use the characterization of a symmetric positive definite matrix via its Schur complement. We have :

$$\Sigma_{\kappa+1} = \underbrace{\begin{pmatrix} \Sigma & \Sigma-D & \dots & \Sigma-D \\ \Sigma-D & \Sigma & \dots & \Sigma-D \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma-D & \Sigma-D & \dots & \Sigma \end{pmatrix}}_{\kappa+2 \text{ blocks}} \succ 0$$

$$\Leftrightarrow \underbrace{\begin{pmatrix} \Sigma & \dots & \Sigma-D \\ \vdots & \ddots & \vdots \\ \Sigma-D & \dots & \Sigma \end{pmatrix}}_{\kappa+1 \text{ blocks}}$$

$$- \begin{pmatrix} \Sigma-D \\ \vdots \\ \Sigma-D \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \Sigma-D & \dots & \Sigma-D \end{pmatrix} \succ 0$$

$$\Leftrightarrow \underbrace{\begin{pmatrix} C & \dots & C-D \\ \vdots & \ddots & \vdots \\ C-D & \dots & C \end{pmatrix}}_{\kappa+1 \text{ blocks}} \succ 0, \ C \text{ defined below}$$

$$\Leftrightarrow \frac{1+\kappa}{\kappa} C - D \succ 0, \quad \text{by induction as } C \succ 0$$

$$\Leftrightarrow \frac{2+\kappa}{1+\kappa} D - D\Sigma^{-1}D \succ 0$$

$$\Leftrightarrow \begin{pmatrix} \Sigma & D \\ D & \frac{2+\kappa}{1+\kappa} D \end{pmatrix} \succ 0$$

$$\Leftrightarrow \Sigma - D(\frac{2+\kappa}{1+\kappa} D)^{-1} D \succ 0$$

$$\text{as } \Sigma \succ 0 \text{ and this is a Schur complement}$$

$$\Leftrightarrow \frac{2+\kappa}{1+\kappa} \Sigma - D \succ 0$$

Hence the recursive step and we conclude the proof. We have

$$C = \Sigma - (\Sigma - D)\Sigma^{-1}(\Sigma - D) = 2D - D\Sigma^{-1}D \succ 0$$

given that $\Sigma \succ 0$ so C is the Schur complement of :

$$\begin{pmatrix} \Sigma & \Sigma - D \\ \Sigma - D & \Sigma \end{pmatrix} \succ 0$$

Objective Function for Entropy Construction In addition to this, we need to formulate the objective function for the entropy construction. The entropy of a multivariate Gaussian has a simple closed formula.

$$H(\boldsymbol{X}^0, \boldsymbol{X}^1, \dots, \boldsymbol{X}^\kappa) = \frac{1}{2} \log \det(2\pi e \Sigma_\kappa)$$

We rearrange the expression of $\det(\Sigma_\kappa)$ to show that minimizing $-\log(\det(2\pi e \Sigma_\kappa))$ is equivalent to minimizing

$$-\log\det(\frac{\kappa+1}{\kappa}\Sigma - D(s)) - \kappa \sum_{i=1}^{d} \log(s_i)$$

(We showed in the main text that minimizing the entropy in a Gaussian setting is equivalent to minimizing this log-determinant). In order to do so, it suffices to show by induction that the following holds for all $\kappa \geq 1$:

$$\det(\Sigma_\kappa) \propto \det(D)^\kappa \det(\frac{\kappa+1}{\kappa}\Sigma - D)$$

where the multiplicative constant is a real number depending only on κ. We first show this for $\kappa = 1$.

$$\begin{aligned} \det(\Sigma_1) &= \det \begin{pmatrix} \Sigma & \Sigma - D(s) \\ \Sigma - D(s) & \Sigma \end{pmatrix} \\ &= \det(\Sigma) \det \left(\Sigma - (\Sigma - D(s))\Sigma^{-1}(\Sigma - D(s)) \right) \\ &= \det(\Sigma) \det \left(2D(s) - D(s)\Sigma^{-1}D(s) \right) \\ &= \det \begin{pmatrix} \Sigma & D(s) \\ D(s) & 2D(s) \end{pmatrix} \\ &= \det(2\Sigma D(s) - D(s)D(s)) \\ &= \det(2\Sigma - D(s)) \prod_{i=1}^{d} s_i \end{aligned}$$

Suppose the result holds for a given $\kappa \geq 1$. We use the notation $|A| = \det(A)$. We have:

$$|\Sigma_{\kappa+1}| = \underbrace{\begin{vmatrix} \Sigma & \Sigma-D & \dots & \Sigma-D \\ \Sigma-D & \Sigma & \dots & \Sigma-D \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma-D & \Sigma-D & \dots & \Sigma \end{vmatrix}}_{\kappa+2 \text{ blocks}}$$

$$= |\Sigma| \Bigg| \underbrace{\begin{pmatrix} \Sigma & \dots & \Sigma-D \\ \vdots & \ddots & \vdots \\ \Sigma-D & \dots & \Sigma \end{pmatrix}}_{\kappa+1 \text{ blocks}}$$

$$- \begin{pmatrix} \Sigma-D \\ \vdots \\ \Sigma-D \end{pmatrix} \Sigma^{-1} \begin{pmatrix} \Sigma-D & \dots & \Sigma-D \end{pmatrix} \Bigg|$$

$$=|\Sigma|\underbrace{\begin{vmatrix} C & \dots & C-D \\ \vdots & \ddots & \vdots \\ C-D & \dots & C \end{vmatrix}}_{\kappa+1 \text{ blocks}}$$

$$\propto|\Sigma||D|^{\kappa}\Big|\frac{1+\kappa}{\kappa}C-D\Big| \qquad \text{by induction}$$

$$\propto|\Sigma||D|^{\kappa}\Big|\frac{2+2\kappa}{\kappa}D-\frac{1+\kappa}{\kappa}D\Sigma^{-1}D-D\Big|$$

$$\propto|\Sigma||D|^{\kappa+1}\Big|\frac{2+\kappa}{\kappa}I-\frac{1+\kappa}{\kappa}\Sigma^{-1}D\Big|$$

$$\propto|D|^{\kappa+1}\Big|\frac{2+\kappa}{\kappa+1}\Sigma-D\Big|$$

Hence the result, where C is the same as before. We used the following two formulae to compute determinants of block matrices:

- If A is invertible, then
$$\det\begin{pmatrix} A & B \\ C & D \end{pmatrix}=\det(A)\det(D-CA^{-1}B)$$
- If C and D commute and all the blocks are square matrices, then $\det\begin{pmatrix} A & B \\ C & D \end{pmatrix}=\det(AD-BC)$

Lower Bound for Diagonal Terms in Entropy Construction For the entropy construction, in order to give a lower bound for the s_i, we derive an expression for the solution of the minimization problem. Without loss of generality, fix $j\in\{1,\dots,d\}$ so that we compute the partial derivative with respect to s_j. Denote $R(s)=\frac{\kappa+1}{\kappa}\Sigma-D(s)$. Using Jacobi's formula for the derivative of a determinant, we get:

$$\frac{\mathrm{d}}{\mathrm{d}s_j}\big(|R(s)|\prod_{i=1}^{d}s_i^{\kappa}\big)$$

$$=\Big(\frac{\mathrm{d}}{\mathrm{d}s_j}|R(s)|\Big)\prod_{i=1}^{d}s_i^{\kappa}+|R(s)|\big(\prod_{i\neq j}^{d}s_i\big)s_j^{\kappa-1}$$

$$=|R(s)|tr\Big(R(s)^{-1}\frac{\mathrm{d}R(s)}{\mathrm{d}s_j}\Big)\prod_{i=1}^{d}s_i+|R(s)|s_j^{\kappa-1}\prod_{i\neq j}^{d}s_i$$

$$=(s_j^{\kappa-1}-s_jR(s)^{-1}_{jj})\Big(|R(s)|\prod_{i\neq j}^{d}s_i\Big)$$

given that $\frac{\mathrm{d}}{\mathrm{d}s_j}R(s)=-\frac{\mathrm{d}}{\mathrm{d}s_j}D(s)=-\mathbb{I}_{jj}$ where $\mathbb{I}_{kl}$ is a matrix where the only non-zero term equal to one is in position (kl). Therefore $tr\Big(R(s)^{-1}\frac{\mathrm{d}R(s)}{\mathrm{d}s_j}\Big)=-R(s)^{-1}_{jj}$. Setting this expression to 0 we get that the solution of the convex optimization problem satisfies

$$\frac{1}{s_j^{\kappa-2}}=R(s)^{-1}_{jj} \qquad \forall j\in\{1,\dots,d\}$$

Now we can write the diagonal term in the inverse matrix as a quotient between two determinants: $R(s)^{-1}_{jj}=\frac{M_j(s)}{\det(R(s))}$ where $M_j(s)$ is the principal minor of $R(s)$ when removing the jth row and column. As both $M_j(s)$ and $\det R(s)$ can be written as a product of eigenvalues, the Cauchy interlacing theorem gives the following lower bound:

$$\lambda_{min}(R(s))\leq\min_{j\in\{1,\dots,d\}}s_j^{\kappa-2}$$

where $\lambda_{min}(R(s))$ is the smallest (positive) eigenvalue of $R(s)$. □

A.2 General Multi-knockoff Sampling Based on SCIP

We can also generalize to the multi-knockoff setting a universal (although possibly intractable) knockoff sampling algorithm introduced in Candès et al. (2018): the Sequential Conditional Independent Pairs (SCIP). Fix $\kappa\geq1$ the number of multi-knockoffs to sample (so that SCIP corresponds to $\kappa=1$). We iterate for $1\leq i\leq d$ over the features, at each step sampling κ knockoffs for the ith feature, independently one of another, from the conditional distribution of the original feature given all the available variables sampled so far. It is important to notice that, whenever SCIP is tractable due to the particular structure of a given initial feature distribution (as for Hidden Markov Models), this generalization to multi-knockoffs will also be tractable given that increasing the number of multi-knockoffs does not alter the conditional dependencies between knockoffs and original features. We formulate this in Algorithm 2 and prove that the resulting samples satisfy exchangeability.

Algorithm 2: Sequential Conditional Independent Multi-knockoffs

```
for $1\leq i\leq d$ do
    for $1\leq k\leq\kappa$ do
        Sample $X_i^k\sim\mathcal{L}(X_i^0|\boldsymbol{X}_{-i}^0,X_{1:i-1}^{1:\kappa})$
    end
end
return $X_{1:d}^{1:\kappa}$
```

Proof. We need to prove the following equality in distribution, using the notations of Definition 3.1:

$$[\boldsymbol{X}^0,\boldsymbol{X}^1,\dots\boldsymbol{X}^{\kappa}]_{swap(\sigma)}\overset{d}{=}[\boldsymbol{X}^0,\boldsymbol{X}^1,\dots\boldsymbol{X}^{\kappa}]$$

We follow the same proof as in Candès et al. (2018), where we have the following induction hypothesis:

Induction Hypothesis: After i steps, we have

$$[\boldsymbol{X}^0,\boldsymbol{X}^1_{1:i},\dots\boldsymbol{X}^{\kappa}_{1:i}]_{swap(\sigma)}\overset{d}{=}[\boldsymbol{X}^0,\boldsymbol{X}^1_{1:i},\dots\boldsymbol{X}^{\kappa}_{1:i}]$$

where now $\sigma=(\sigma_j)_{1\leq j\leq i}$ with arbitrary permutations σ_j over $\{0,\dots,\kappa\}$. After the first step the equality holds for $i=1$ given that all X_1^k have the same conditional distribution and are independent one of another. Now, if the hypothesis holds at step $i-1$, then at step i we have that the joint distribution of $[\boldsymbol{X}^0, \boldsymbol{X}^1_{1:i}, \dots \boldsymbol{X}^\kappa_{1:i}]$ can be decomposed as a product of conditional distributions given the sampling procedure so that we have:

$$\mathcal{L}(\boldsymbol{X}^0, \boldsymbol{X}^1_{1:i}, \dots \boldsymbol{X}^\kappa_{1:i}) = \frac{\prod_{k=0}^{\kappa} \mathcal{L}(X_i^k | \boldsymbol{X}^0_{-i}, X^{1:\kappa}_{1:i-1})}{\mathcal{L}(\boldsymbol{X}^0_{-i}, X^{1:\kappa}_{1:i-1})}$$

Now, by induction hypothesis, the expression in the denominator satisfies the extended exchangeability for the $i-1$ first dimensions (we marginalize out over X_i^0 which doesn't matter as at step $i-1$ the permutations σ_j are over $j \leq i-1$). And so are the terms in the numerator, as again we permute only elements among the first $i-1$ dimensions. And, because of the conditional independent sampling, the numerator expression is also exchangeable for the ith dimension. In conclusion, $\mathcal{L}(\boldsymbol{X}^0, \boldsymbol{X}^1_{1:i}, \dots \boldsymbol{X}^\kappa_{1:i})$ is exchangeable for the first i dimensions, hence concluding the proof. □

B PROOFS

B.1 Proof of Lemma 3.1

Proof. Given that the $swap(\sigma)$ operation is the concatenation of the action of each permutation σ_i onto $(X_i^0, \dots, X_i^\kappa)$ and that we can write σ_i as the composition of transpositions, we see that it is enough to show the result for a simple transposition of two features (original or multi-knockoff) corresponding to a null dimension. This leads us directly to the proof of Lemma 3.2 in Candès et al. (2018), where the difference is that we add all the extra multi-knockoffs in the conditioning set. □

B.2 Proof of Lemma 3.2

Proof. Consider any collection $(\sigma_i)_{i \in \mathcal{H}_0}$ of permutations σ_i on the set $\{0, \dots, \kappa\}$, and for $i \notin \mathcal{H}_0$, set $\sigma_i = ()$ the identity permutation. In order to prove the result we need to show the following equality in distribution:

$$\Big([\sigma_i(\kappa_i)]_{1\leq i\leq d}, [(T_i^{(k)})_{0\leq k\leq\kappa}]_{1\leq i\leq d}\Big) \overset{d}{=} \Big([\kappa_i]_{1\leq i\leq d}, [(T_i^{(k)})_{0\leq k\leq\kappa}]_{1\leq i\leq d}\Big)$$

Define $\hat{T}_i^k = T_i^{\sigma_i(k)}$ for every $i \in \{1, \dots, d\}$ and $k \in \{0, \dots, \kappa\}$. Using the notation for the extended swap this is equivalent to $\hat{\boldsymbol{T}} = \boldsymbol{T}_{swap(\sigma)}$, where for each null index $i \in \mathcal{H}_0$ the ith features of T and its knockoffs have been permuted according to σ_i (and the non-null remained at their place). By construction, $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X}, Y)$ is a function of $\boldsymbol{X}$ and Y which associates to each feature in $\boldsymbol{X}$ a "score" for its importance (for simplicity here we will denote by $\boldsymbol{X}$ the whole concatenated vector of $[\boldsymbol{X}^0, \boldsymbol{X}^1, \dots \boldsymbol{X}^\kappa]$). The choice of such function is restricted so that $\boldsymbol{T}_{swap(\sigma)} = \boldsymbol{T}(\boldsymbol{X}_{swap(\sigma)}, Y)$. By the multi-knockoff exchangeability property, and our specific choice of σ that does not permute non-null features, we also have $(\boldsymbol{X}_{swap(\sigma)}, Y) \overset{d}{=} (\boldsymbol{X}, Y)$. This in turn implies:

$$\hat{\boldsymbol{T}} \overset{d}{=} \boldsymbol{T}$$

Also, given that the permutation is done feature-wise, the feature-wise ordered importance scores remain the same.

$$[(\hat{T}_i^{(k)})_{0\leq k\leq\kappa}]_{1\leq i\leq d} = [(T_i^{(k)})_{0\leq k\leq\kappa}]_{1\leq i\leq d}$$

We now prove the equality in distribution (where we have an abusive notation for representing set probabilities):

$$\begin{aligned}
&\mathbb{P}([\kappa_i]_{1\leq i\leq d}, [(T_i^{(k)})_{0\leq k\leq\kappa}]_{1\leq i\leq d}) \\
&= \mathbb{P}([T_i^{\kappa_i} = T_i^{(0)}]_{1\leq i\leq d}, [(T_i^{(k)})_{0\leq k\leq\kappa}]_{1\leq i\leq d}) \\
&= \mathbb{P}([\hat{T}_i^{\kappa_i} = \hat{T}_i^{(0)}]_{1\leq i\leq d}, [(\hat{T}_i^{(k)})_{0\leq k\leq\kappa}]_{1\leq i\leq d}) \\
&= \mathbb{P}([T_i^{\sigma_i(\kappa_i)} = T_i^{(0)}]_{1\leq i\leq d}, [(T_i^{(k)})_{0\leq k\leq\kappa}]_{1\leq i\leq d}) \\
&= \mathbb{P}([\sigma_i(\kappa_i)]_{1\leq i\leq d},\ [(T_i^{(k)})_{0\leq k\leq\kappa}]_{1\leq i\leq d})
\end{aligned}$$

The second equality is due to the equality in distribution between $\boldsymbol{T}$ and $\hat{\boldsymbol{T}}$, and the third equality makes use of the fact that for any $i \in \{1, \dots, d\}$ the order statistics of $(\hat{T}_i^0, \dots, \hat{T}_i^\kappa)$ and $(T_i^0, \dots, T_i^\kappa)$ are the same. The statement about our variables τ_i holds because they are functions of the feature-wise ordered importance scores. □

B.3 Proof of Proposition 3.3

Proof. The random variables κ_i allow us to construct one-bit p-values as in Barber et al. (2015), while the τ_i can be used to determine the ordering in which we sort those p-values, given that conditionally on $(\tau_i)_{1\leq i\leq d}$, we have $(\kappa_i)_{i\in\mathcal{H}_0}$ i.i.d. uniform over $\{0, \dots, \kappa\}$, independent of $(\kappa_i)_{i\notin\mathcal{H}_0}$. We can therefore permute the dimension indices based on $(\tau_i)_i$ so that $\tau_1 \geq \tau_2 \geq \dots \geq \tau_d \geq 0$, and still define the following random variables with the desired properties. We expect that our ordering based on $(\tau_i)_i$ will tend to place non-nulls at the beginning. Set for $1 \leq i \leq d$:

$$p_i = \begin{cases} \frac{1}{\kappa+1}, & \kappa_i = 0 \\ 1, & \kappa_i \geq 1 \end{cases}$$

The distributional results for $(\kappa_i)_{i\in\mathcal{H}_0}$ imply that the null $(p_i)_{i\in\mathcal{H}_0}$ are also i.i.d., independent of the non-null $(p_i)_{i\notin\mathcal{H}_0}$ and the $(\tau_i)_{1\leq i\leq d}$ and have the following distribution:

$$\begin{cases} \mathbb{P}(p_i = \frac{1}{\kappa+1}) = \frac{1}{\kappa+1} \\ \mathbb{P}(p_i = 1) = \frac{\kappa}{\kappa+1} \end{cases}$$

In particular, null p_i satisfy $p_i \overset{d}{\geq} \mathcal{U}([0,1])$. Fix a target FDR level $q \in (0,1)$, and a constant $c \in (0,1)$. Following Barber et al. (2015), define the Selective SeqStep+

threshold:

$$\hat{k} = \max\left\{1 \le k \le d,\ \frac{1+\#\{i \le k : p_i > c\}}{\#\{i \le k : p_i \le c\} \vee 1} \le \frac{1-c}{c} q\right\}$$

Then according to Theorem 3 in Barber et al. (2015), the procedure that selects the features $\mathcal{S} = \{i \le \hat{k},\ p_i \le c\}$, controls for FDR at level q. For the particular choice of $c = \frac{1}{\kappa+1}$, we have:

$$\begin{aligned}\hat{k} &= \max\left\{1 \le k \le d, \frac{1+\#\{i \le k : p_i > \frac{1}{\kappa+1}\}}{\#\{i \le k : p_i \le \frac{1}{\kappa+1}\} \vee 1} \le \kappa q\right\}\\ &= \max\left\{1 \le k \le d, \frac{1+\#\{i \le k : \kappa_i \ge 1\}}{\#\{i \le k : \kappa_i = 0\} \vee 1} \le \kappa q\right\}\\ &= \max\left\{1 \le k \le d, \frac{\frac{1}{\kappa}+\frac{1}{\kappa}\#\{i : \kappa_i \ge 1,\ \tau_i \ge \tau_k\}}{\#\{i : \kappa_i = 0,\ \tau_i \ge \tau_k\} \vee 1} \le q\right\}\end{aligned}$$

Now, instead of maximizing over k indexing a decreasing sequence $\tau_1 \ge \cdots \ge \tau_d$, one can formulate the problem as minimizing the threshold τ:

$$\tau^* = \min\left\{\tau > 0, \frac{\frac{1}{\kappa}+\frac{1}{\kappa}\#\{1 \le i \le d : \kappa_i \ge 1, \tau_i \ge \tau\}}{\#\{1 \le i \le d : \kappa_i = 0, \tau_i \ge \tau\} \vee 1} \le q\right\}$$

The selection set is then defined as:

$$\hat{\mathcal{S}} = \{i \in \{1, \ldots, d\},\ \kappa_i = 0,\ \tau_i \ge \hat{\tau}\}$$

□

We notice that the main role of τ_i is to determine an ordering sequence of the p-values for the Adaptive SeqStep+ procedure. Any function of the ordered statistics $(T_i^{(k)})_{0 \le k \le \kappa}$ gives valid statistics that can be used to order the p-values, given that the distributional restrictions will still be satisfied. A rich literature covers this topic (Lei and Fithian, 2018; Lei et al., 2017; Ignatiadis et al., 2016), and could be applied to multi-knockoff based p-values.

B.4 Intuition for Choice of Kappa and Tau

We illustrate the particular choice of (κ_i) and (τ_i) from a geometric point of view. For the single knockoffs, one can pair the importance statistics of each original feature and its knockoff $(T_i, \tilde{T}_i)$ and plot such pairs as points in a plane $\mathbb{R}_+^2$. We then have a geometric view of the threshold selection. Consider the parallel lines given by the equations $y = x + t$ and $y = x - t$, partitioning the plane into 3 sections. The terms $\#\{j : W_j \le -t\}$ and $\#\{j : W_j \ge t\}$ in the FDP estimate

$$\widehat{FDP}_{KN+} = \frac{1+\#\{j : W_j \le -t\}}{\#\{j : W_j \ge t\} \vee 1}$$

are obtained by counting the number of points $(T_i, \tilde{T}_i)$ in the section above $y = x + t$ (that is, $y \ge x + t$) and below $y = x - t$ (that is, $y \le x - t$). For $t = 0$, the two lines collapse and $\mathbb{R}_+^2$ is partitioned by the line $y = x$.

The same setting can happen in higher dimensions, where we partition the space $\mathbb{R}^d$ into d cones given by $C_i = \{x \in \mathbb{R}_+^d, x_i = \max_j x_j\}$. Our method for choosing a threshold for multi-knockoffs proceeds as before: for a given $t > 0$, we count the number of points $(T_i^0, T_i^1, \ldots, T_i^\kappa) \in \mathbb{R}^{\kappa+1}$ in each translated cone $C_{it} = \{x \in \mathbb{R}_+^d, x_i \ge t + \max_{j \ne i} x_j\}$ and compare the counts in C_{0t} corresponding to the original feature to the average over those in C_{it}. We then find the minimum t subject to some constraint. Reformulating this gives our variables κ_i and τ_i.

C SUPPLEMENT ON SIMULATIONS

C.1 Comparison Between Distributions of Diagonal Construction Methods

We run another simulation where we increase the dimension of the samples. We plot again the distribution of the logarithm of the diagonal terms for the three construction methods in Figure 5. As we increase the dimension, we observe that the distributions are shifted towards more negative values, indicating that the diagonal coefficients constructed tend to be smaller. This is particularly the case for the equicorrelated construction. The SDP construction generates an even higher proportion of almost-zero diagonal terms as we increase the dimension. Also, increasing the level of correlation has also an impact on the distribution of the diagonal terms similar to what we observe by increasing the dimension.

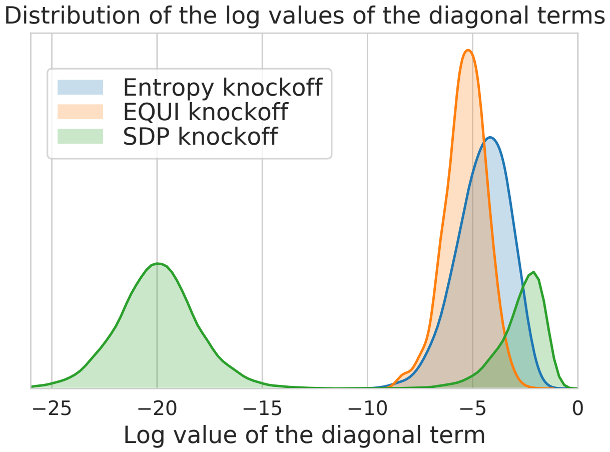

Figure 5: **Comparison Between Diagonal Matrix Construction Methods - Increased Dimension to 400**

C.2 Measuring Stability of the Set of SDP-based Undiscoverable Features with Jaccard Similarity

For a given correlation matrix, we generate samples from a centered multivariate Gaussian. Based on the estimated correlation matrix from these samples, we run the SDP construction to get the matrix D, and identify the set of undiscoverable features. By repeatedly doing this, we obtain multiple sets of undiscoverable features. In Figure 6 we plot the averaged Jaccard similarity over all pairs of such sets, as a function of the sample size (and repeat the whole procedure 50 times to estimate the variance of our results). Even though the similarity increases with the sample size, it remains very low. Furthermore, the similarity decreases with the dimension, so in high-dimensional problems where $d >> N$ then the SDP construction method is very unstable, and has a very high proportion of undiscoverable features as suggested in Figure 5. Reproducing findings becomes then very hard in such settings if we use SDP knockoffs.

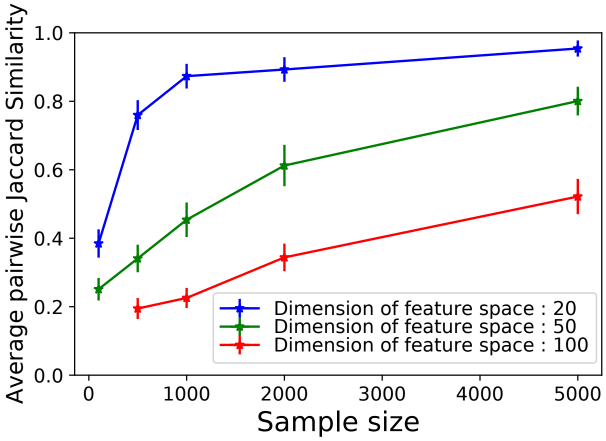

Figure 6: **Average Pairwise Jaccard Similarity for Multiple Runs of SDP Method**

C.3 Comparing Power Between SDP and Entropy Knockoffs

We show an extreme example of the drastic improvement in power brought by entropy knockoffs over SDP knockoffs. We generate a dataset $(\boldsymbol{X}, Y)$ where we specify the distribution of the feature set such that we can predict which diagonal coefficients will be set to 0 in the SDP method, and thus construct the response Y such that the non-null features are undiscoverable. We choose a particular covariance structure that conveniently allows for explicit expressions of the diagonal terms in each method, though the results apply more generally as shown in Figure 1.

We sample $\boldsymbol{X} \sim \mathcal{N}(0, \Sigma)$ as a multivariate centered Gaussian random variable, where $\Sigma \in \mathbb{R}^{3d\times 3d}$ is a covariance matrix defined as a block-diagonal matrix:

$$\Sigma = \underbrace{\begin{pmatrix} A & 0 & 0 & \dots & 0 & 0 \\ 0 & A & 0 & \dots & 0 & 0 \\ 0 & 0 & A & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A & 0 \\ 0 & 0 & 0 & \dots & 0 & A \end{pmatrix}}_{d \text{ blocks}}$$

where $A = \begin{pmatrix} 1 & a & 0 \\ a & 1 & a \\ 0 & a & 1 \end{pmatrix}$ for some $a \geq 0$.

SDP and entropy methods output a diagonal matrix $D = sI_{3d}$ such that $s \in \mathbb{R}^{3d}$ is the concatenation d times the sequence $(s_1, s_2, s_1) \in \mathbb{R}^3$, which corresponds to the output of the corresponding method on the matrix A. We can derive an explicit formula for s_1, s_2 as functions of a for both the SDP and entropy methods, which we denote $(s_1^{SDP}(a), s_2^{SDP}(a))$ and $(s_1^{entr}(a), s_2^{entr}(a))$. We plot such curves in Figures 7, which show that for a wide range of values of a, $s_2^{SDP}(a)$ is exactly equal to 0, whereas the diagonal terms of the entropy method stay always positive. Notice that in this particular setting the maximal value that a can take is $\frac{1}{\sqrt{2}}$, otherwise the convex optimization problem has an empty feasible set.

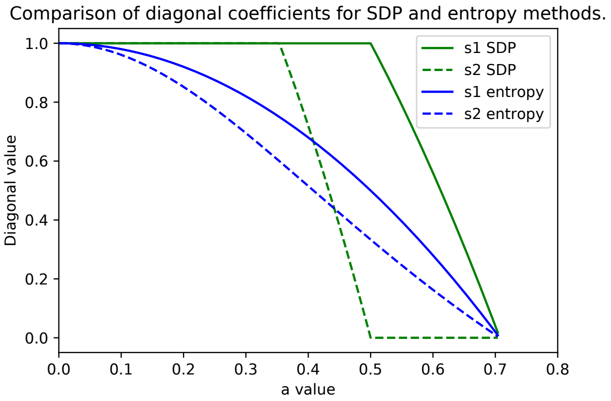

Figure 7: **Comparison of Diagonal Values for SDP and Entropy Methods** For values of a in the range $[\frac{1}{2}, \frac{1}{\sqrt{2}})$ the value $s_2^{SDP}(a)$ is exactly equal to 0: the optimization objective favors setting s_2^{SDP} to 0 in order to maximize s_1^{SDP}. Entropy knockoffs do not suffer from this issue.

This phenomenon becomes worse as we increase the number of simultaneously correlated features, we refer again to Figures 1 and 5.

We now generate a large number of samples so that the estimated empirical correlation matrix is very close to the real one (so that sample size is not a factor when comparing SDP and entropy methods). We then sample a response vector Y such that the non-null features correspond to the dimensions associated to the s_2 diagonal terms (i.e. the non-null features are given by $\mathcal{H}_0 = \{3i+2,\ 0 \leq i \leq (d-1)\}$). The non-null features are therefore undiscoverable under the SDP construction, whereas entropy knockoffs are still able to select the non-nulls. The results of simulating the whole procedure are clear: SDP has zero power, and entropy knockoffs have full power. Of course, this is an extreme situation designed to accentuate this behavior. Still, across the multiple simulations done in this paper, entropy knockoffs consistently had higher power than SDP knockoffs.

C.4 Generating the Synthetic Response for the Real Genome Dataset

We collect data from the 1000 Genomes Project (Consortium et al., 2015), and obtain around 2000 individual samples for 27 distinct segments of chromosome 19 containing an average of 50 SNPs per segment. We filter out SNPs that are extremely correlated (above 0.95), and generate for each of those 27 segments a random subset that will correspond to the causal SNPs. We then generate the response accordingly and use a logistic regression to obtain importance scores. For the top correlation method, we select the top k correlated features, where k is chosen as the number of rejections that multi-knockoffs make, so that we have a fair comparison between methods.

One important caveat that explains why sometimes the averaged FDP is above the target is that with real data, it is crucial to accurately estimate the feature distribution. In these simulations, we approximate the 0/1/2 matrix of SNPs by a Gaussian distribution, where we need to estimate the covariance based on the data. Such inaccurate approximation causes the average FDP to exceed the target sometimes. However, the knockoff procedure is robust to mis-estimations of the feature distribution (Barber et al., 2018), so that we can expect FDR control at an inflated level. Our FDR results are therefore satisfactory, and the comparison is stark with the top correlation method that catastrophically fails to control FDR.