Maximum Entropy Distributions on Unitary Orbits

Polynomial Capacity: Theory, Applications, Generalizations

Jonathan Leake

Technische Universität Berlin

February 4th, 2020

Outline

Outline

Relative entropy

Given a base measure ν on a support set $\mathcal{S} \subset \mathbb{R}^n$, let $\mu := \phi \cdot \nu$ be another measure given via a function $\phi : \mathcal{S} \to \mathbb{R}_+$. Define **relative entropy**:

$$\mathsf{D}_{\mathsf{KL}}(\mu \| \nu) := \int_{\mathcal{S}} \phi(\boldsymbol{x}) \log \phi(\boldsymbol{x}) d\nu(\boldsymbol{x}).$$

Discrete case: Let p be a polynomial associated to μ supported on $\mathcal{S} \subset \mathbb{Z}_+^n$ via $p(\boldsymbol{x}) = \sum_{\boldsymbol{v} \in \mathcal{S}} \mu(\boldsymbol{v}) \cdot \boldsymbol{x}^{\boldsymbol{v}}$. When ν is uniform on $\mathcal{S}$ we have

$$\mathsf{D}_{\mathsf{KL}}(\mu \| \nu) = \sum_{\boldsymbol{v} \in \mathcal{S}} \left[\frac{\mu(\boldsymbol{v})}{1/|\mathcal{S}|} \log \frac{\mu(\boldsymbol{v})}{1/|\mathcal{S}|} \right] \frac{1}{|\mathcal{S}|} = \log |\mathcal{S}| - \mathcal{H}(\mu),$$

where $\mathcal{H}(\mu)$ is the **entropy** of μ. **Properties** for fixed $\mathcal{S}$:

- When μ is a probability distribution, $\mathcal{H}(\mu) \geq 0$.
- When μ, ν probability distributions, $\mathsf{D}_{\mathsf{KL}}(\mu \| \nu) \geq 0$.
- $\mathcal{H}(\mu)$ is maximized for $\mu =$ the uniform distribution.
- $\mathsf{D}_{\mathsf{KL}}(\mu \| \nu)$ is minimized for $\mu = \phi \cdot \nu$ when $\phi \equiv 1$.

Maximum entropy distributions

Last slide: For distributions ν and $\mu := \phi \cdot \nu$ on $\mathcal{S}$:

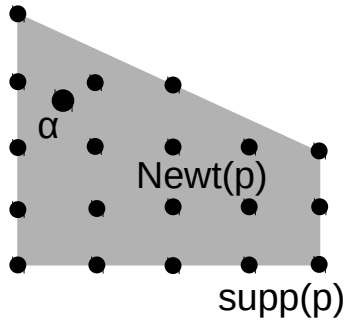
$$\mathsf{D}_{\mathsf{KL}}(\mu \| \nu) = \int \phi(\boldsymbol{x}) \log \phi(\boldsymbol{x}) d\nu(\boldsymbol{x}).$$

Discrete case: Also $\mathcal{H}(\nu) := -\sum_{\boldsymbol{v} \in \mathcal{S}} \nu(\boldsymbol{v}) \log \nu(\boldsymbol{v})$.

Maximum entropy distribution: Fix $\boldsymbol{\alpha}$ in $\mathsf{hull}(\mathcal{S})$ for discrete $\mathcal{S}$.

$$\mathcal{H}_{\mathsf{opt}} := \sup_{\substack{\mathsf{supp}(\nu)=\mathcal{S} \\ \mathbb{E}[\nu]=\boldsymbol{\alpha}}} \mathcal{H}(\nu) \quad \text{with} \quad \nu_{\mathsf{opt}} := \arg\sup .$$

Discrete picture for $p \sim \nu$:



Optimum always of the form $\nu_{\mathsf{opt}}(\boldsymbol{v}) = e^{\langle \boldsymbol{y}, \boldsymbol{v} \rangle}$ for some $\boldsymbol{y}$ (via capacity).

Minimum relative entropy distributions

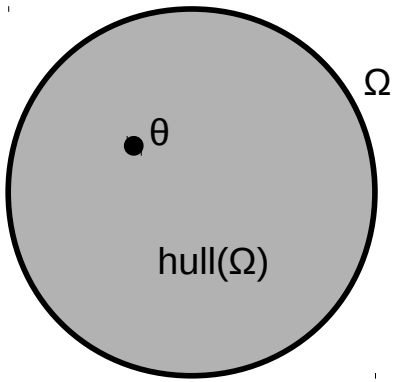
Previous slide: For distributions ν and $\mu := \phi \cdot \nu$ on $\mathcal{S}$:

$$\mathsf{D}_{\mathsf{KL}}(\mu \| \nu) = \int \phi(\boldsymbol{x}) \log \phi(\boldsymbol{x}) d\nu(\boldsymbol{x}).$$

Minimum relative entropy ("maximum entropy") distribution:
Fix $\boldsymbol{\theta}$ in hull$(\mathcal{S})$ for any $\mathcal{S}$, and fix some base measure ν on $\mathcal{S}$.

$$\mathcal{H}_{\mathsf{opt}} := \inf_{\substack{\mu=\phi\cdot\nu \\ \mathbb{E}[\mu]=\boldsymbol{\theta}}} \mathsf{D}_{\mathsf{KL}}(\mu \| \nu) \quad \text{with} \quad \mu_{\mathsf{opt}} := \arg\inf .$$

Continuous picture where $\Omega = \mathcal{S}$:



Optimum always of the form $\mu_{\mathsf{opt}} = e^{\langle \boldsymbol{y}, \boldsymbol{x} \rangle} \cdot \nu$ for some $\boldsymbol{y}$ (via capacity).

Relation to capacity

Last slide: Fix θ in hull($\mathcal{S}$) for any $\mathcal{S}$, and base measure ν on $\mathcal{S}$.

$$\mathcal{H}_{\text{opt}} := \inf_{\substack{\mu=\phi\cdot\nu \\ \mathbb{E}[\mu]=\theta}} \mathrm{D}_{\mathrm{KL}}(\mu\|\nu) \quad \text{with} \quad \mu_{\text{opt}} := \arg\inf .$$

This min relative entropy program has the following **dual formulation**:

$$\inf_{\substack{\mu=\phi\cdot\nu \\ \mathbb{E}[\mu]=\theta}} \mathrm{D}_{\mathrm{KL}}(\mu\|\nu) = -\inf_{\boldsymbol{y}\in\mathbb{R}^n} \left[\log \int e^{\langle \boldsymbol{y},\boldsymbol{v}\rangle} d\nu(\boldsymbol{v}) - \langle \boldsymbol{y},\theta\rangle\right] = -\log \mathrm{Cap}_\theta(p),$$

where $p(\boldsymbol{x}) = \int \boldsymbol{x}^{\boldsymbol{v}} d\nu(\boldsymbol{v})$ and $\mathrm{Cap}_\theta(p) = \inf_{\boldsymbol{x}>0} \dfrac{\int \boldsymbol{x}^{\boldsymbol{v}} d\nu(\boldsymbol{v})}{\boldsymbol{x}^\theta}$.

Strong duality: $\mu_{\text{opt}} = e^{\langle \boldsymbol{y}_{\text{opt}},\boldsymbol{v}\rangle}\cdot\nu = \boldsymbol{x}_{\text{opt}}^{\boldsymbol{v}}\cdot\nu$ where $\boldsymbol{x}_{\text{opt}}$ is the optimum input for $\mathrm{Cap}_\theta(p) \implies$ Max entropy optimization equivalent to $\mathrm{Cap}_\theta(p)$.

- Infinite-dimensional max entropy vs n-dimensional capacity.
- Ellipsoid method for capacity via evaluation and gradient oracles.

Outline

Unitary orbits

Unitary group: Set of $n \times n$ complex matrices U such that $U^* = U^{-1}$.

Hermitian matrix: An $n \times n$ complex matrix H such that $H = H^*$.

- $H = \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^*$ for $\lambda_1 \geq \cdots \geq \lambda_n \in \mathbb{R}$ and $\boldsymbol{v}_i$ orthonormal.
- **Majorization property:** $\sum_{i=1}^{k} \lambda_i \geq \sum_{i=1}^{k} h_{i,i}^{\downarrow}$ for all k (= for n).
- For any unitary U, matrix UHU^* is Hermitian with same eigenvalues.

Unitary "adjoint" orbits: Orbit of H under conjugation action.

- **Equivalent:** Set of Hermitian matrices with specified eigenvalues.
- **"Adjoint":** Orbits under action of unitary group on its Lie algebra.

E.g.: Orbit of $D = \mathrm{diag}(1, 0, \ldots, 0)$ is set of rank-one projection matrices.

Convex hull of this orbit is the set of trace-1 positive semidefinite matrices.

Majorization: Diagonal entries of any UDU^* gives a simplex element.

Moment map

Last slide: $D = \mathrm{diag}(1, 0, \ldots, 0)$, orbit of D:

$$\mathcal{O}_D = \{UDU^* : U \text{ unitary}\} = \{H : \mathrm{eig}(H) = (1, 0, \ldots, 0)\}.$$

Fact: The map diag : $H \mapsto (h_{11}, \ldots, h_{nn})$ is such that:

$$\mathrm{diag}(\mathcal{O}_D) = \Delta_n := \left\{ \boldsymbol{x} \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1 \right\} = \mathrm{hull}(\{\sigma \cdot \mathrm{eig}(D) : \sigma \in S_n\}).$$

Proof: For any $\boldsymbol{x} \in \Delta_n$, pick $H = \sqrt{\boldsymbol{x}} \cdot \sqrt{\boldsymbol{x}}^\top$ where $\sqrt{\cdot}$ is entrywise.

More general: For any D, we have that $\mathrm{diag}(\mathcal{O}_D)$ is a convex polytope.
Proof idea: The majorization inequalities translate to:

$$\sum_{i=1}^k \lambda_i \geq \sum_{i=1}^k h_{\sigma(i),\sigma(i)} \quad \text{for all } k, \sigma \qquad \text{and} \qquad \sum_{i=1}^n \lambda_i = \mathrm{tr}(H) = \sum_{i=1}^n h_{i,i}.$$

These cut out the correct polytope, called the **moment polytope**.
We call the map diag the **moment map**. (From torus action, generally.)
E.g.: $(1, 0, \ldots, 0) \to$ simplex, $(n, n-1, \ldots, 1) \to$ permutohedron.

Measures on orbits

Given real diagonal D, consider unitary conjugation orbit $\mathcal{O}_D$.

Unitary Haar measure: "Uniform" measure on unitary group $U(n)$.

- Unique unitarily invariant probability measure on $U(n)$:

$$\int_{U(n)} f(U)dU = \int_{U(n)} f(VU)dU = \int_{U(n)} f(UV)dU.$$

- Easily sampled by iteratively constructing random orthonormal basis.
- Pushforward through the map $U \mapsto UDU^*$ gives conjugation-invariant **Haar measure** on the orbit $\mathcal{O}_D$: $\int_{\mathcal{O}_D} f(H)dH = \int_{\mathcal{O}_D} f(UHU^*)dH$.

Looking ahead: This "uniform" Haar measure is a great candidate for the base measure on $\mathcal{O}_D$ for the maximum entropy program on $\mathcal{O}_D$.

Further: Push forward this measure through diag: "pushes forward" max entropy distributions. **What's the base measure on the polytope?**

Last week: For $D = \mathrm{diag}(1, 0, \ldots, 0)$, this measure is uniform on Δ_n.

Moment polytope measure for unitary orbits

Last slide: For $D = \mathrm{diag}(1, 0, \ldots, 0)$, the pushforward measure through diag is uniform on Δ_n:

$$\int_{\mathcal{O}_D} f(\mathrm{diag}(H))dH = \int_{\Delta_n} f(\boldsymbol{x})d\boldsymbol{x}.$$

E.g.: Let $Y = \mathrm{diag}(\boldsymbol{y})$ for some $\boldsymbol{y} \in \mathbb{R}^n$. Then:

$$\int_{\mathcal{O}_D} e^{\langle Y,H\rangle} dH = \int_{\Delta_n} e^{\langle \boldsymbol{y},\boldsymbol{x}\rangle} d\boldsymbol{x} \quad \Longrightarrow \quad \text{max entropy correspondence.}$$

Problem: Does not hold for general D. **Duistermaat-Heckman:** Pushforward is a piecewise polynomial density times uniform measure.

Solution: Higher-dimensional polytope and another "moment-like" map for which the pushforward is uniform: **Gelfand-Tsetlin (GT) polytope**.

More general question: Can we evaluate the Duistermaat-Heckman measure? Piecewise polynomial denisty very generally for toric action on symplectic manifold, density is always log-concave, etc...

Gelfand-Tsetlin polytope

Hermitian matrix H with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$: let $\mathcal{R}$ denote the map $H \mapsto (R_{ij})_{1 \leq i \leq j \leq n}$ where $R_{i,j}$ is the i^{th} largest eigenvalue of the $j \times j$ submatrix in the top left corner of H. **E.g.:** $R_{i,n} = \lambda_i$ and $R_{1,1} = H_{1,1}$.

Cauchy interlacing theorem: $R_{1,j} \geq R_{1,j-1} \geq R_{2,j} \geq R_{2,j-1} \geq R_{3,j} \geq \cdots$

Rayleigh triangle: Organizing the $R_{i,j}$ in a natural way:

$$
\begin{array}{ccccccccc}
R_{1,n} & & R_{2,n} & & R_{3,n} & & R_{4,n} & \cdots & R_{n,n} \\
& R_{1,n-1} & & R_{2,n-1} & & R_{3,n-1} & \cdots & R_{n-1,n-1} & \\
& & R_{1,n-2} & & R_{2,n-2} & \cdots & R_{n-2,n-2} & & \\
& & & \ddots & & & & &
\end{array}
$$

Cauchy interlacing theorem $\implies$ the Cauchy inequalities cut out the **Gelfand-Tsetlin polytope** $\mathrm{GT}(\boldsymbol{\lambda})$ of all triangles with top row $\boldsymbol{\lambda}$.

Fun fact: The image $\mathcal{R}(\mathcal{O}_D)$ is $\mathrm{GT}(\boldsymbol{\lambda})$, and the pushforward of the invariant measure dH on $\mathcal{O}_D$ is the uniform measure on $\mathrm{GT}(\boldsymbol{\lambda})$.

Outline

Ellipsoid method

Ellipsoid method for convex optimization: Iterations of the algorithm:

1. Have an ellipsoid which contains the optimum input.
2. Compute the gradient at the center of the ellipsoid: Determines which half of the ellipsoid contains the optimum.
3. Construct a new ellipsoid containing that half.

At the end: Small ellipsoid containing the optimum $\implies$ approximation.

Need: (1) Oracle for the gradient, and (2) starting ellipsoid.

(There are more details to this, but this is roughly the idea. We'll discuss this a bit further, but see [L-Vishnoi '20], [Singh-Vishnoi '15] for details.)

Recall the convex optimization (capacity) problem:

$$-\inf_{\boldsymbol{y}\in\mathbb{R}^n}\left[\log\int e^{\langle \boldsymbol{y},\boldsymbol{v}\rangle}d\nu(\boldsymbol{v})-\langle \boldsymbol{y},\boldsymbol{\theta}\rangle\right]=-\inf_{\boldsymbol{y}\in\mathbb{R}^n}\left[\log\int e^{\langle \boldsymbol{y},\boldsymbol{v}-\boldsymbol{\theta}\rangle}d\nu(\boldsymbol{v})\right].$$

Need: Oracle for $\nabla\log\int e^{\langle \boldsymbol{y},\boldsymbol{v}\rangle}d\nu(\boldsymbol{v})$ and "bounding box" for $\boldsymbol{y}_{\text{opt}}$.

Applying ellipsoid method to unitary orbit capacity

Unitary orbits: $\mathcal{O}_D$ for real diagonal D with invariant Haar measure dH:

$$\text{Optimize} \quad -\inf_Y \left[\log \int_{\mathcal{O}_D} e^{\langle Y, H-\Theta\rangle} dH\right] \quad \text{for } \Theta \in \text{hull}(\mathcal{O}_D).$$

Last slide: Need oracle for $\nabla \log \int e^{\langle Y,H\rangle} dH$ and bound on Y_{opt}.

First: dH is a pushforward measure of dU through $U \mapsto UDU^*$:

$$\int_{\mathcal{O}_D} e^{\langle Y,H\rangle} dH = \int_{U(n)} e^{\langle Y,UDU^*\rangle} dU.$$

HCIZ formula: For $\text{eig}(D) = (\lambda_1, \ldots, \lambda_n)$ and $\text{eig}(Y) = (y_1, \ldots, y_n)$:

$$\int_{U(n)} e^{\langle Y,UDU^*\rangle} dU = \frac{\prod_{i=1}^{n-1} i!}{\prod_{i<j}(\lambda_i - \lambda_j)\prod_{i<j}(y_i - y_j)} \cdot \det\left(\left[e^{\lambda_i y_j}\right]_{i,j=1}^{n}\right).$$

Apply gradient to get the oracle. **Also:** L'Hopital to handle multiplicity.

Next: Bound on Y_{opt}. Unitary invariance $\implies$ "balanced" measure.
Depends on distance δ of Θ from boundary: $\|Y_{\text{opt}}\| \leq \text{poly}(n, \log\|D\|, \frac{1}{\delta})$.

Alternative method for computing max entropy

Previous slides: Invariant measure dH on $\mathcal{O}_D$ corresponds to uniform measure on a polytope (either through diag of Ralyeigh map $\mathcal{R}$).

Rank-one case: Polytope is simplex and $Y = \mathrm{diag}(\boldsymbol{y})$ gives

$$\int_{\mathcal{O}_D} e^{\langle Y,H\rangle}\, dH = \int_{\Delta_n} e^{\langle \boldsymbol{y},\boldsymbol{x}\rangle}\, d\boldsymbol{x}.$$

Therefore: Computing integrals / max entropy over $\mathcal{O}_D$ is equivalent to integrals / max entropy over Δ_n. Utilize algos which compute integrals over convex polytopes (e.g. [Lovász-Vempala '06]).

Similarly: For general D and $Y = \mathrm{diag}(\boldsymbol{y})$, we have

$$\int_{\mathcal{O}_D} e^{\langle Y,H\rangle}\, dH = \int_{\mathrm{GT}(\boldsymbol{\lambda})} e^{\langle \boldsymbol{y},\mathrm{type}(R)\rangle}\, dR.$$

Here: R is a Rayleigh triangle, and $\mathrm{type}(R)$ outputs the vector of "diagonal entries" of R (i.e., $h_{kk} = \sum_{i=1}^{k} R_{i,k} - \sum_{i=1}^{k-1} R_{i,k-1}$).

Again: Utilize algos which compute integrals over convex polytopes.

Outline

Sampling from the associated polytope

Recall: The uniform distribution (dH) on $\mathcal{O}_D$ is easy to sample from.

Idea: Choose random unitary and conjugate D to get sample UDU^*.

Problem: What about max entropy distributions $e^{\langle Y,H\rangle}dH$?

Idea: Sample from the polytope, then "bring back" to the orbit.

Previous slides: For $D_0 = (1, 0, \ldots, 0)$, general D, $Y = \mathrm{diag}(\boldsymbol{y})$:

$$\int_{\mathcal{O}_{D_0}} e^{\langle Y,H\rangle}dH = \int_{\Delta_n} e^{\langle \boldsymbol{y},\boldsymbol{x}\rangle}d\boldsymbol{x}, \quad \int_{\mathcal{O}_D} e^{\langle Y,H\rangle}dH = \int_{\mathrm{GT}(\boldsymbol{\lambda})} e^{\langle \boldsymbol{y},\mathrm{type}(R)\rangle}dR.$$

Ways to sample: General methods for sampling from log-concave (even log-linear) distributions from convex polytopes.

Or: One can get a good handle on the entrywise CDFs conditioned on various other coordinates: always determinants here. **Sample?**

Next problem: How to go back to the orbit $\mathcal{O}_D$?

Returning to the orbit: Rank-one case

Last slide: Sample $\boldsymbol{x}$ from Δ_n according to $e^{\langle \boldsymbol{y}, \boldsymbol{x} \rangle} d\boldsymbol{x}$.

Next question: How do we convert $\boldsymbol{x}$ into a sample from $\mathcal{O}_D$ (the orbit of rank-one projections) according to $e^{\langle Y, H \rangle} dH = e^{\langle Y, \boldsymbol{v}\boldsymbol{v}^* \rangle} d\boldsymbol{v}$?

Let's compute the fiber of a given $\boldsymbol{x} \in \Delta_n$:

$$\operatorname{diag}^{-1}(\boldsymbol{x}) = \{\boldsymbol{v}\boldsymbol{v}^* : \operatorname{diag}(\boldsymbol{v}\boldsymbol{v}^*) = \boldsymbol{x}\} = \mathbb{T}^n \cdot \sqrt{\boldsymbol{x}}\sqrt{\boldsymbol{x}}^\top.$$

That is: $\boldsymbol{v} = (e^{i\theta_1}\sqrt{x_1}, \ldots, e^{i\theta_n}\sqrt{x_n})$. Since Y is diagonal, we then have

$$e^{\langle Y, \boldsymbol{v}\boldsymbol{v}^* \rangle} = e^{\langle \operatorname{diag}(e^{-i\boldsymbol{\theta}}) Y \operatorname{diag}(e^{i\boldsymbol{\theta}}), \sqrt{\boldsymbol{x}}\sqrt{\boldsymbol{x}}^\top \rangle} = e^{\langle Y, \sqrt{\boldsymbol{x}}\sqrt{\boldsymbol{x}}^\top \rangle}$$

That is: The max entropy distribution is uniform on fibers.

Therefore: To construct a rank-1 projection sample from a simplex sample, we just need to uniformly sample from $\mathbb{T}^n$ and multiply by $\sqrt{\boldsymbol{x}}$.

Returning to the orbit: General case

Now: Sample Rayleigh triangle R from $\mathrm{GT}(\boldsymbol{\lambda})$ according to $e^{\langle \boldsymbol{y}, \mathrm{type}(R) \rangle} dR$.

Next: How to convert R into a sample from $\mathcal{O}_D$ according to $e^{\langle Y, H \rangle} dH$?

Problem: Fibers of R are more complicated. (Recall that R holds the eigenvalues of all leading principal submatrices).

Solution: Induct on the rows of R. Assume we have $k \times k$ sample according to the bottom k rows, and construct a $(k+1) \times (k+1)$ matrix using the next $((k+1)^{\mathrm{st}})$ row up.

$$
\begin{array}{ccccccccc}
R_{1,n} & & R_{2,n} & & R_{3,n} & & R_{4,n} & \cdots & R_{n,n} \\
& R_{1,n-1} & & R_{2,n-1} & & R_{3,n-1} & \cdots & R_{n-1,n-1} & \\
& & R_{1,n-2} & & R_{2,n-2} & \cdots & R_{n-2,n-2} & & \\
& & & \ddots & & & & &
\end{array}
$$

That is: Given $k \times k$ matrix X_0, we want to sample $\begin{bmatrix} X_0 & \boldsymbol{u} \\ \boldsymbol{u}^* & c \end{bmatrix}$.

Key: Max entropy distributions are uniform on these "partial fibers".

Returning to the orbit: Partial fibers

Task: Given $k \times k$ Hermitian X_0 with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k$, want to sample uniformly random matrix $\begin{bmatrix} X_0 & \boldsymbol{u} \\ \boldsymbol{u}^* & c \end{bmatrix}$ with eigenvalues $\gamma_1 \geq \cdots \geq \gamma_{k+1}$.

For simplicity: We assume entries of $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are all distinct.

1. Diagonalize X_0 using $\begin{bmatrix} U & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{bmatrix}$ to get $\begin{bmatrix} D_0 & \boldsymbol{w} \\ \boldsymbol{w}^* & c \end{bmatrix}$ where $\operatorname{diag}(D_0) = \boldsymbol{\lambda}$.
2. Compute $c = \sum_{i=1}^{k+1} \gamma_i - \sum_{i=1}^{k} \lambda_i$.
3. **Hardest part:** Sample $\boldsymbol{w}$. Recall that $\boldsymbol{\lambda}, \boldsymbol{\gamma}$ interlace and note that:

$$\prod_{i=1}^{k+1}(z-\gamma_i) = \det\left(zI - \begin{bmatrix} D_0 & \boldsymbol{w} \\ \boldsymbol{w}^* & c \end{bmatrix}\right)$$

$$= (z-c)\prod_{i=1}^{k}(z-\lambda_i) - \sum_{i=1}^{k}|w_i|^2 \prod_{j\neq i}(z-\lambda_j).$$

   Plug in $z = \lambda_j$, rearrange: $|w_j|^2 = \frac{\prod_{i=1}^{k+1}(\lambda_j - \gamma_i)}{\prod_{i\neq j}(\lambda_j - \lambda_i)}$, sample $e^{i\theta}|w_j|$.

4. Apply $\begin{bmatrix} U & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{bmatrix}^*$ to convert D_0 back to X_0.