Equilibrating low-rank approximations with Gaussian priors & High-performance finite DPP sampling via mirror-image Cholesky

Jack Poulson

Google Research

ELSI Conference, August 2018
Overview

1. Equilibrating low-rank approximations with Gaussian priors

2. High-performance finite DPP sampling via mirror-image Cholesky
Motivation for analyzing equilibration

Recommender systems and language models often involve low-rank approximations of a large, sparse matrix $A$, e.g., a local minimum of:

$$
\mathcal{L}(X, Y) = \frac{1}{2} \| W \circ (A - X Y^*) \|_F^2 + \frac{\lambda}{2} \left( \| X \|_F^2 + \| Y \|_F^2 \right),
$$

where $W$ is a weighting matrix (often a function of $A$).\(^1\)

This is Maximum Likelihood inference with $(X Y^*)_{i,j} \sim \mathcal{N}(A_{i,j}, W_{i,j}^{-2})$ and priors $X_{i,j}, Y_{i,j} \sim \mathcal{N}(0, 1/\lambda)$.\(^2\)

One can find an approximate local minimum via a few iterations of Weighted Alternating Least Squares.\(^3\)

A colleague (Steffen Rendle) observed that results for his model satisfied $X^* X = Y^* Y$. How do we prove (and exploit) this property?

---

\(^1\)See, for example, [Hu et al.-2008] Collaborative filtering for implicit feedback datasets


\(^3\)http://www.tensorflow.org/api_docs/python/tf/contrib/factorization/WALSModel
Motivation for analyzing equilibration

Recommender systems and language models often involve low-rank approximations of a large, sparse matrix \( A \), e.g., a local minimum of:

\[
\mathcal{L}(X, Y) = \frac{1}{2} \| W \circ (A - XY^*) \|_F^2 + \frac{\lambda}{2} \left( \| X \|_F^2 + \| Y \|_F^2 \right),
\]

where \( W \) is a weighting matrix (often a function of \( A \)).

This is Maximum Likelihood inference with \((XY^*)_{i,j} \sim \mathcal{N}(A_{i,j}, W_{i,j}^{-2})\) and priors \(X_{i,j}, Y_{i,j} \sim \mathcal{N}(0, 1/\lambda)\).

One can find an approximate local minimum via a few iterations of Weighted Alternating Least Squares.

A colleague (Steffen Rendle) observed that results for his model satisfied \(X^*X = Y^*Y\). How do we prove (and exploit) this property?

---

1See, for example, [Hu et al.-2008] Collaborative filtering for implicit feedback datasets
3http://www.tensorflow.org/api_docs/python/tf/contrib/factorization/WALSModel
Motivation for analyzing equilibration

Recommender systems and language models often involve low-rank approximations of a large, sparse matrix $A$, e.g., a local minimum of:

$$
\mathcal{L}(X, Y) = \frac{1}{2} \| W \circ (A - XY^*) \|^2_F + \frac{\lambda}{2} \left( \| X \|^2_F + \| Y \|^2_F \right),
$$

where $W$ is a weighting matrix (often a function of $A$).\(^1\)

This is Maximum Likelihood inference with $(XY^*)_i,j \sim \mathcal{N}(A_{i,j}, W_{i,j}^{-2})$ and priors $X_{i,j}, Y_{i,j} \sim \mathcal{N}(0, 1/\lambda)$.\(^2\)

One can find an approximate local minimum via a few iterations of Weighted Alternating Least Squares.\(^3\)

A colleague (Steffen Rendle) observed that results for his model satisfied $X^*X = Y^*Y$. How do we prove (and exploit) this property?

---

\(^1\) See, for example, [Hu et al.-2008] Collaborative filtering for implicit feedback datasets


\(^3\) [http://www.tensorflow.org/api_docs/python/tf/contrib/factorization/WALSModel](http://www.tensorflow.org/api_docs/python/tf/contrib/factorization/WALSModel)
Motivation for analyzing equilibration

Recommender systems and language models often involve low-rank approximations of a large, sparse matrix $A$, e.g., a local minimum of:

$$\mathcal{L}(X, Y) = \frac{1}{2} \| W \circ (A - XY^*) \|_F^2 + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right),$$

where $W$ is a weighting matrix (often a function of $A$).\(^1\)

This is Maximum Likelihood inference with $(XY^*)_{i,j} \sim \mathcal{N}(A_{i,j}, W_{i,j}^{-2})$ and priors $X_{i,j}, Y_{i,j} \sim \mathcal{N}(0, 1/\lambda)$.\(^2\)

One can find an approximate local minimum via a few iterations of Weighted Alternating Least Squares.\(^3\)

A colleague (Steffen Rendle) observed that results for his model satisfied $X^*X = Y^*Y$. How do we prove (and exploit) this property?

---

\(^1\)See, for example, [Hu et al.-2008] Collaborative filtering for implicit feedback datasets


\(^3\)http://www.tensorflow.org/api_docs/python/tf/contrib/factorization/WALSModel
Why the Gramians are equivalent [1/3]

**Definition 1.** Given $S \in \text{Sym}(n, \mathbb{R})$, we will use the shorthand $P(S)$ for the linear operator $P(S) : \text{Sym}(n, \mathbb{R}) \to \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$.

**Definition 2.** The geometric mean of $A, B \in S^n_{++}$ is $A \# B = B \# A = P(A^{1/2})(P(A^{-1/2})B)^{1/2}$.

**Proposition 1.** For any $A, B \in S^n_{++}$, there is a unique $S \in S^n_{++}$ such that $P(S)A = B$.

**Proof.** For existence, put $S = A^{-1} \# B$. For uniqueness, if $P(S)A = P(T)A$, then $X^*AX = A$, with $X = T^{-1}S$. Then the spectral decomposition $(S^{1/2}T^{-1}S^{1/2})(S^{1/2}Z) = (S^{1/2}Z)\Lambda$ implies $XZ = Z\Lambda$, $\Lambda \succ 0$. And $Z^*AZ = Z^*(X^*AX)Z = \Lambda Z^*AZ\Lambda$, so $\Lambda = I$ and $T = S$. \(\square\)

**Definition 3.** The Nesterov-Todd scaling point of $A, B \in S^n_{++}$ is $P(S^{1/2})A = P(S^{-1/2})B$, where $S = A^{-1} \# B$.

---


Why the Gramians are equivalent [1/3]

**Definition 1.** Given $S \in \text{Sym}(n, \mathbb{R})$, we will use the shorthand $P(S)$ for the linear operator $P(S) : \text{Sym}(n, \mathbb{R}) \rightarrow \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$.

**Definition 2.** The geometric mean of $A, B \in S^{n}_{++}$ is $A \# B = B \# A = P(A^{1/2})(P(A^{-1/2})B)^{1/2}$.

**Proposition 1.** For any $A, B \in S^{n}_{++}$, there is a unique $S \in S^{n}_{++}$ such that $P(S)A = B$.

**Proof.** For existence, put $S = A^{-1} \# B$.

For uniqueness, if $P(S)A = P(T)A$, then $X^*AX = A$, with $X = T^{-1}S$. Then the spectral decomposition $(S^{1/2}T^{-1}S^{1/2})(S^{1/2}Z) = (S^{1/2}Z)\Lambda$ implies $XZ = Z\Lambda, \Lambda \succ 0$. And $Z^*AZ = Z^*(X^*AX)Z = \Lambda Z^*AZ\Lambda$, so $\Lambda = I$ and $T = S$. \qed

**Definition 3.** The Nesterov-Todd scaling point of $A, B \in S^{n}_{++}$ is $P(S^{1/2})A = P(S^{-1/2})B$, where $S = A^{-1} \# B$.

---


5 [Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones
Why the Gramians are equivalent [1/3]

**Definition 1.** Given \( S \in \text{Sym}(n, \mathbb{R}) \), we will use the shorthand \( P(S) \) for the linear operator \( P(S) : \text{Sym}(n, \mathbb{R}) \to \text{Sym}(n, \mathbb{R}) \) via \( P(S)A = SAS \).

**Definition 2.** The geometric mean of \( A, B \in S_++^n \) is \( A \# B = B \# A = P(A^{1/2})(P(A^{-1/2})B)^{1/2} \).

**Proposition 1.** For any \( A, B \in S_++^n \), there is a unique \( S \in S_++^n \) such that \( P(S)A = B \).

**Proof.** For existence, put \( S = A^{-1} \# B \).

For uniqueness, if \( P(S)A = P(T)A \), then \( X^*AX = A \), with \( X = T^{-1}S \). Then the spectral decomposition \( (S^{1/2}T^{-1}S^{1/2})(S^{1/2}Z) = (S^{1/2}Z)\Lambda \) implies \( XZ = Z\Lambda, \Lambda \succ 0 \). And \( Z^*AZ = Z^*(X^*AX)Z = \Lambda Z^*AZ\Lambda \), so \( \Lambda = I \) and \( T = S \). \( \square \)

**Definition 3.** The Nesterov-Todd scaling point of \( A, B \in S_++^n \) is \( P(S^{1/2})A = P(S^{-1/2})B \), where \( S = A^{-1} \# B \).

---


\(^5\) [Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones
Why the Gramians are equivalent [1/3]

**Definition 1.** Given $S \in \text{Sym}(n, \mathbb{R})$, we will use the shorthand $P(S)$ for the linear operator $P(S) : \text{Sym}(n, \mathbb{R}) \to \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$.

**Definition 2.** The *geometric mean* of $A, B \in S^n_{++}$ is $A \# B = B \# A = P(A^{1/2})(P(A^{-1/2})B)^{1/2}$.

**Proposition 1.** For any $A, B \in S^n_{++}$, there is a unique $S \in S^n_{++}$ such that $P(S)A = B$.

**Proof.** For existence, put $S = A^{-1} \# B$.

For uniqueness, if $P(S)A = P(T)A$, then $X^*AX = A$, with $X = T^{-1}S$. Then the spectral decomposition $(S^{1/2}T^{-1}S^{1/2})(S^{1/2}Z) = (S^{1/2}Z)\Lambda$ implies $XZ = Z\Lambda$, $\Lambda \succ 0$. And $Z^*AZ = Z^*(X^*AX)Z = \Lambda Z^*AZ\Lambda$, so $\Lambda = I$ and $T = S$. \qed

**Definition 3.** The *Nesterov-Todd scaling point* of $A, B \in S^n_{++}$ is $P(S^{1/2})A = P(S^{-1/2})B$, where $S = A^{-1} \# B$.

---


5 [Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones
Why the Gramians are equivalent [1/3]

**Definition 1.** Given $S \in \text{Sym}(n, \mathbb{R})$, we will use the shorthand $P(S)$ for the linear operator $P(S) : \text{Sym}(n, \mathbb{R}) \to \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$.

**Definition 2.** The **geometric mean** of $A, B \in S^n_{++}$ is $A \# B = B \# A = P(A^{1/2})(P(A^{-1/2})B)^{1/2}$.

**Proposition 1.** For any $A, B \in S^n_{++}$, there is a unique $S \in S^n_{++}$ such that $P(S)A = B$.

**Proof.** For existence, put $S = A^{-1} \# B$.

For uniqueness, if $P(S)A = P(T)A$, then $X^*AX = A$, with $X = T^{-1}S$. Then the spectral decomposition $(S^{1/2}T^{-1}S^{1/2})(S^{1/2}Z) = (S^{1/2}Z)\Lambda$ implies $XZ = Z\Lambda$, $\Lambda \succ 0$. And $Z^*AZ = Z^*(X^*AX)Z = \Lambda Z^*AZ \Lambda$, so $\Lambda = I$ and $T = S$. □

**Definition 3.** The **Nesterov-Todd scaling point** of $A, B \in S^n_{++}$ is $P(S^{1/2})A = P(S^{-1/2})B$, where $S = A^{-1} \# B$.

---


$^5$[Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones
Definition 1. Given $S \in \text{Sym}(n, \mathbb{R})$, we will use the shorthand $P(S)$ for the linear operator $P(S) : \text{Sym}(n, \mathbb{R}) \rightarrow \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$.

Definition 2. The \textbf{geometric mean} of $A, B \in S_{++}^n$ is $A \# B = B \# A = P(A^{1/2})(P(A^{-1/2})B)^{1/2}$.

Proposition 1. For any $A, B \in S_{++}^n$, there is a unique $S \in S_{++}^n$ such that $P(S)A = B$.

Proof. For existence, put $S = A^{-1} \# B$.

For uniqueness, if $P(S)A = P(T)A$, then $X^*AX = A$, with $X = T^{-1}S$. Then the spectral decomposition $(S^{1/2}T^{-1}S^{1/2})(S^{1/2}Z) = (S^{1/2}Z)\Lambda$ implies $XZ = Z\Lambda$, $\Lambda \succ 0$. And $Z^*AZ = Z^*(X^*AX)Z = \Lambda Z^*AZ\Lambda$, so $\Lambda = I$ and $T = S$. \qed

Definition 3. The \textbf{Nesterov-Todd scaling point} of $A, B \in S_{++}^n$ is $P(S^{1/2})A = P(S^{-1/2})B$, where $S = A^{-1} \# B$.


\[\text{[Nesterov/Todd-1998]}\] Primal-Dual Interior Point Methods for self-scaled cones
Lemma 4 (P.). Given \((X, Y) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}, S \in S_{++}^n\) minimizes \(f : S_{++}^n \rightarrow \mathbb{R}_+\), where

\[ f(S) = \|XS\|_F^2 + \|YS^{-1}\|_F^2, \]

iff \(P(S)(X^*X) = P(S^{-1})(Y^*Y)\). And, if \(X\) and \(Y\) have full column rank, then \(S = ((X^*X)^{-1} \# (Y^*Y))^{1/2}\) is the unique minimizer.

Proof. Decompose \(f\) as \(g \circ h\), where \(h : S_{++}^n \rightarrow S_{++}^n\) via \(h(S) = S^2\) and \(g : S_{++}^n \rightarrow \mathbb{R}_+\) via \(g(T) = \langle X^*X, T \rangle + \langle Y^*Y, T^{-1} \rangle\).

Then \(h\) is a diffeomorphism and \(dg_T : (T_S^n \cong \text{Sym}(n, \mathbb{R})) \rightarrow (T_{g(T)} \mathbb{R} \cong \mathbb{R})\) via \(dg_T(dT) = \langle X^*X - T^{-1}Y^*YT^{-1},dT \rangle\).

So \(S \in S_{++}^n\) is a critical point of \(f\) iff \(df_S = dg_{S^2} \circ dh_S = 0\) iff \(X^*X - S^{-2}Y^*YS^{-2} = 0\).
Lemma 4 (P.). Given \((X, Y) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}\), \(S \in S_{++}^n\) minimizes 
\[ f : S_{++}^n \to \mathbb{R}_+, \text{ where} \]
\[ f(S) = \|XS\|_F^2 + \|YS^{-1}\|_F^2, \]
iff \(P(S)(X^*X) = P(S^{-1})(Y^*Y)\). And, if \(X\) and \(Y\) have full column rank, then 
\(S = ((X^*X)^{-1} \# (Y^*Y))^{1/2}\) is the unique minimizer.

**Proof.** Decompose \(f\) as \(g \circ h\), where 
\[ h : S_{++}^n \to S_{++}^n \text{ via } h(S) = S^2 \]
and 
\[ g : S_{++}^n \to \mathbb{R}_+ \text{ via } g(T) = \langle X^*X, T \rangle + \langle Y^*Y, T^{-1} \rangle. \]

Then \(h\) is a diffeomorphism and 
\[ dg_T : (T_T S_{++}^n \cong \text{Sym}(n, \mathbb{R})) \to (T_{g(T)} \mathbb{R} \cong \mathbb{R}) \]
via 
\[ dg_T(dT) = \langle X^*X - T^{-1}Y^*YT^{-1}, dT \rangle. \]

So \(S \in S_{++}^n\) is a critical point of \(f\) iff 
\[ df_S = dg_{S^2} \circ dh_S = 0 \text{ iff } \]
\[ X^*X - S^{-2}Y^*YS^{-2} = 0. \]
Lemma 4 (P.). Given \((X, Y) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}\), \(S \in S_{++}^n\) minimizes \(f : S_{++}^n \to \mathbb{R}_+\), where

\[
f(S) = \|XS\|_F^2 + \|YS^{-1}\|_F^2,
\]

iff \(P(S)(X^*X) = P(S^{-1})(Y^*Y)\). And, if \(X\) and \(Y\) have full column rank, then \(S = ((X^*X)^{-1} \# (Y^*Y))^{1/2}\) is the unique minimizer.

Proof. Decompose \(f\) as \(g \circ h\), where \(h : S_{++}^n \to S_{++}^n\) via \(h(S) = S^2\) and \(g : S_{++}^n \to \mathbb{R}_+\) via \(g(T) = \langle X^*X, T \rangle + \langle Y^*Y, T^{-1} \rangle\).

Then \(h\) is a diffeomorphism and \(dg_T : (T_T S_{++}^n \cong \text{Sym}(n, \mathbb{R})) \to (T_{g(T)}\mathbb{R} \cong \mathbb{R})\) via \(dg_T(dT) = \langle X^*X - T^{-1}Y^*YT^{-1}, dT \rangle\).

So \(S \in S_{++}^n\) is a critical point of \(f\) iff \(df_S = dg_{S^2} \circ dh_S = 0\) iff \(X^*X - S^{-2}Y^*YS^{-2} = 0\). 

Lemma 4 (P.). Given \((X, Y) \in \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r}\), \(S \in S_{++}^n\) minimizes \(f : S_{++}^n \to \mathbb{R}_+\), where

\[
f(S) = \|XS\|_F^2 + \|YS^{-1}\|_F^2,
\]

iff \(P(S)(X^*X) = P(S^{-1})(Y^*Y)\). And, if \(X\) and \(Y\) have full column rank, then \(S = ((X^*X)^{-1} \# (Y^*Y))^{1/2}\) is the unique minimizer.

Proof. Decompose \(f\) as \(g \circ h\), where \(h : S_{++}^n \to S_{++}^n\) via \(h(S) = S^2\) and \(g : S_{++}^n \to \mathbb{R}_+\) via \(g(T) = \langle X^*X, T \rangle + \langle Y^*Y, T^{-1} \rangle\).

Then \(h\) is a diffeomorphism and \(dg_T : (T_T S_{++}^n \cong \text{Sym}(n, \mathbb{R})) \to (T_{g(T)}\mathbb{R} \cong \mathbb{R})\) via \(dg_T(dT) = \langle X^*X - T^{-1}Y^*YT^{-1}, dT \rangle\).

So \(S \in S_{++}^n\) is a critical point of \(f\) iff \(df_S = dg_{S^2} \circ dh_S = 0\) iff \(X^*X - S^{-2}Y^*YS^{-2} = 0\). \[\square\]
Why the Gramians are equivalent [3/3]

**Theorem 5 (P.).** If \( \ell : \mathbb{R}^{m \times n} \to \mathbb{R} \) is continuous, the local minima of 
\[
\mathcal{L} : \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \to \mathbb{R},
\]
where
\[
\mathcal{L}(X, Y) = \ell(XY^*) + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right),
\]
satisfy \( X^*X = Y^*Y \). And, given any candidate \((X, Y)\), the *equilibration*, 
\((XS^{1/2}, YS^{-1/2})\), where \( S = (X^*X)^{-1} \# (Y^*Y) \), minimizes the regularization 
while preserving the input to \( \ell \).

**Proof.** Given \((X, Y)\), \( \ell(XY^*) \) is invariant under any transformation 
\((X, Y) \mapsto (XZ, YZ^*) \) where \( Z \in GL(n, \mathbb{R}) \). 
Thus, any local minimum must satisfy
\[
\|X\|_F^2 + \|Y\|_F^2 = \min_{Z \in GL(n, \mathbb{R})} \{ \|XZ\|_F^2 + \|YZ^*\|_F^2 \}
\]
\[
= \min_{S \in S^n_{++}} \{ \|XS\|_F^2 + \|YS^{-1}\|_F^2 \},
\]
where we exploited the polar decomposition \( Z = SQ \), \( Q \) unitary. The result 
then follows from our lemma. \( \square \)
Why the Gramians are equivalent [3/3]

Theorem 5 (P.). If \( \ell : \mathbb{R}^{m \times n} \rightarrow \mathbb{R} \) is continuous, the local minima of 
\[ L : \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \rightarrow \mathbb{R}, \]
where
\[ L(X, Y) = \ell(XY^*) + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right), \]
satisfy \( X^*X = Y^*Y \). And, given any candidate \((X, Y)\), the equilibration, \((XS^{1/2}, YS^{-1/2})\), where \( S = (X^*X)^{-1} \# (Y^*Y) \), minimizes the regularization while preserving the input to \( \ell \).

Proof. Given \((X, Y)\), \( \ell(XY^*) \) is invariant under any transformation \((X, Y) \mapsto (XZ, YZ^{-*})\) where \( Z \in GL(n, \mathbb{R}) \).

Thus, any local minimum must satisfy
\[
\|X\|_F^2 + \|Y\|_F^2 = \min_{Z \in GL(n, \mathbb{R})} \left\{ \|XZ\|_F^2 + \|YZ^{-*}\|_F^2 \right\}
\]
\[
= \min_{S \in S_n^{++}} \left\{ \|XS\|_F^2 + \|YS^{-1}\|_F^2 \right\},
\]
where we exploited the polar decomposition \( Z = SQ \), \( Q \) unitary. The result then follows from our lemma. \( \square \)
Why the Gramians are equivalent [3/3]

**Theorem 5 (P.).** If \( \ell : \mathbb{R}^{m \times n} \to \mathbb{R} \) is continuous, the local minima of

\[
\mathcal{L} : \mathbb{R}^{m \times r} \times \mathbb{R}^{n \times r} \to \mathbb{R},
\]

where

\[
\mathcal{L}(X, Y) = \ell(XY^*) + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right),
\]

satisfy \( X^*X = Y^*Y \). And, given any candidate \((X, Y)\), the **equilibration**, \((XS^{1/2}, YS^{-1/2})\), where \( S = (X^*X)^{-1} \# (Y^*Y) \), minimizes the regularization while preserving the input to \( \ell \).

**Proof.** Given \((X, Y)\), \( \ell(XY^*) \) is invariant under any transformation \((X, Y) \mapsto (XZ, YZ^{-*})\) where \( Z \in GL(n, \mathbb{R}) \).

Thus, any local minimum must satisfy

\[
\|X\|_F^2 + \|Y\|_F^2 = \min_{Z \in GL(n, \mathbb{R})} \left\{ \|XZ\|_F^2 + \|YZ^{-*}\|_F^2 \right\}
\]

\[
= \min_{S \in S_n^{++}} \left\{ \|XS\|_F^2 + \|YS^{-1}\|_F^2 \right\},
\]

where we exploited the polar decomposition \( Z = SQ \), \( Q \) unitary. The result then follows from our lemma. \( \square \)
Why the Gramians are equivalent [3/3]

**Theorem 5 (P.).** If \( \ell : \mathbb{R}^{m \times n} \to \mathbb{R} \) is continuous, the local minima of 
\[ L(X, Y) = \ell(XY^*) + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right), \]
satisfy \( X^*X = Y^*Y \). And, given any candidate \((X, Y)\), the *equilibration*, 
\((Xs^{1/2}, Ys^{-1/2})\), where \( S = (X^*X)^{-1} \# (Y^*Y) \), minimizes the regularization while preserving the input to \( \ell \).

**Proof.** Given \((X, Y)\), \( \ell(XY^*) \) is invariant under any transformation 
\((X, Y) \mapsto (XZ, YZ^{-*})\) where \( Z \in GL(n, \mathbb{R}) \).

Thus, any local minimum must satisfy
\[ \|X\|_F^2 + \|Y\|_F^2 = \min_{Z \in GL(n, \mathbb{R})} \{ \|XZ\|_F^2 + \|YZ^{-*}\|_F^2 \} \]
\[ = \min_{S \in S_n^{++}} \{ \|XS\|_F^2 + \|YS^{-1}\|_F^2 \}, \]

where we exploited the polar decomposition \( Z = SQ \), \( Q \) unitary. The result then follows from our lemma. \( \square \)
Equilibrating block coordinate descent

Given

$$\mathcal{L}(X, Y) = \ell(XY^*) + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right),$$

insert an equilibration step between each block coordinate descent step. E.g., if $X$ and $Y$ have full column rank, replace

$$(X, Y) \mapsto (XS^{1/2}, YS^{-1/2}), \quad S = (X^*X)^{-1} \# (Y^*Y),$$

which can be computed in $O((m + n + r)r^2)$ time.

Equilibration is essentially free and keeps the regularization minimized (with the constraint of preserving the loss function input).

If one thinks of $(X^*X, Y^*Y)$ as analogous to a primal/dual pair in an SDP IPM, this is similar to centering the Newton step about the NT point.

Equilibration has a much more pronounced effect for small regularization values.
Equilibrating block coordinate descent

Given

\[ \mathcal{L}(X, Y) = \ell(XY^*) + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right) \]

insert an equilibration step between each block coordinate descent step. E.g., if $X$ and $Y$ have full column rank, replace

\[
(X, Y) \mapsto (XS^{1/2}, YS^{-1/2}), \quad S = (X^*X)^{-1} \# (Y^*Y),
\]

which can be computed in $O((m + n + r)r^2)$ time.

Equilibration is essentially free and keeps the regularization minimized (with the constraint of preserving the loss function input).

If one thinks of $(X^*X, Y^*Y)$ as analogous to a primal/dual pair in an SDP IPM, this is similar to centering the Newton step about the NT point.

Equilibration has a much more pronounced effect for small regularization values.
Equilibrating block coordinate descent

Given
\[ \mathcal{L}(X, Y) = \ell(XY^*) + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right), \]
insert an equilibration step between each block coordinate descent step. E.g., if
\(X\) and \(Y\) have full column rank, replace

\[(X, Y) \mapsto (XS^{1/2}, YS^{-1/2}), \quad S = (X^*X)^{-1} \# (Y^*Y),\]

which can be computed in \(O((m + n + r)r^2)\) time.

Equilibration is essentially free and keeps the regularization minimized (with the constraint of preserving the loss function input).

If one thinks of \((X^*X, Y^*Y)\) as analogous to a primal/dual pair in an SDP IPM, this is similar to centering the Newton step about the NT point.

Equilibration has a much more pronounced effect for small regularization values.
Equilibrating block coordinate descent

Given
\[ \mathcal{L}(X, Y) = \ell(XY^*) + \frac{\lambda}{2} \left( \|X\|_F^2 + \|Y\|_F^2 \right) , \]
insert an equilibration step between each block coordinate descent step. E.g., if
\( X \) and \( Y \) have full column rank, replace
\[ (X, Y) \mapsto (XS^{1/2}, YS^{-1/2}), \quad S = (X^*X)^{-1} \# (Y^*Y) , \]
which can be computed in \( O((m + n + r)r^2) \) time.

Equilibration is essentially free and keeps the regularization minimized (with the
constraint of preserving the loss function input).

If one thinks of \( (X^*X, Y^*Y) \) as analogous to a primal/dual pair in an SDP
IPM, this is similar to centering the Newton step about the NT point.

Equilibration has a much more pronounced effect for small regularization
values.
A trivial example

Consider minimizing \((\alpha - \chi \eta)^2 + \lambda (\chi^2 + \eta^2)\) given \(\alpha = 1, \lambda = 0.001, \chi_0 = \eta_0 = 2\).
Handling ill-conditioned Gramians [1/2]

The Nesterov-Todd equilibration obviously makes assumptions about the invertibility of the Gramians.

Geometrically, $S = A\# B$, when $A, B \in S^n_{++}$, is well-known to be the Euclidean midpoint between $\log(A)$ and $\log(B)$ and the midpoint of the geodesic between $A$ and $B$ when $S^n_{++}$ is equipped with the left-invariant metric $g_X(S, T) = \langle X^{-1}S, X^{-1}T \rangle$.

One could extend the geometric mean to the boundary via:

$$A \# B = \lim_{\epsilon \downarrow 0} (A + \epsilon I) \# (B + \epsilon I).$$

But this extension is discontinuous [Bhatia-2007]: Let

$$A = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 20 & 6 \\ 6 & 2 \end{pmatrix}, X_n = \begin{pmatrix} 1 & 0 \\ 0 & 1/n \end{pmatrix} \to X = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$ 

Then, for $\Phi_n(A) = X_n^*AX_n$, $\Phi_n(A) \# \Phi_n(B) = \Phi_n(A \# B)$.

But sequential continuity is violated:

$$\lim_{n \uparrow \infty} \Phi_n(A) \# \Phi_n(B) = \lim_{n \uparrow \infty} \Phi_n(A \# B) = \Phi(A \# B) = \begin{pmatrix} 8 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\left(\lim_{n \uparrow \infty} \Phi_n(A)\right) \# \left(\lim_{n \uparrow \infty} \Phi_n(B)\right) = \Phi(A) \# \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$
Handling ill-conditioned Gramians [1/2]

The Nesterov-Todd equilibration obviously makes assumptions about the invertibility of the Gramians.

Geometrically, $S = A \# B$, when $A, B \in S_{++}^n$, is well-known to be the Euclidean midpoint between $\log(A)$ and $\log(B)$ and the midpoint of the geodesic between $A$ and $B$ when $S_{++}^n$ is equipped with the left-invariant metric $g_X(S, T) = \langle X^{-1}S, X^{-1}T \rangle$.

One could extend the geometric mean to the boundary via:

$$A \# B = \lim_{\epsilon \downarrow 0} (A + \epsilon I) \# (B + \epsilon I).$$

But this extension is discontinuous [Bhatia-2007]: Let

$$A = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 20 & 6 \\ 6 & 2 \end{pmatrix}, X_n = \begin{pmatrix} 1 & 0 \\ 0 & 1/n \end{pmatrix} \to X = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Then, for $\Phi_n(A) = X_n^* A X_n$, $\Phi_n(A) \# \Phi_n(B) = \Phi_n(A \# B)$.

But sequential continuity is violated:

$$\lim_{n \uparrow \infty} \Phi_n(A) \# \Phi_n(B) = \lim_{n \uparrow \infty} \Phi_n(A \# B) = \Phi(A \# B) = \begin{pmatrix} 8 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\left(\lim_{n \uparrow \infty} \Phi_n(A)\right) \# \left(\lim_{n \uparrow \infty} \Phi_n(B)\right) = \Phi(A) \# \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$
Handling ill-conditioned Gramians [1/2]

The Nesterov-Todd equilibration obviously makes assumptions about the invertibility of the Gramians.

Geometrically, $S = A \# B$, when $A, B \in S^n_{++}$, is well-known to be the Euclidean midpoint between $\log(A)$ and $\log(B)$ and the midpoint of the geodesic between $A$ and $B$ when $S^n_{++}$ is equipped with the left-invariant metric $g_{X}(S, T) = \langle X^{-1}S, X^{-1}T \rangle$.

One could extend the geometric mean to the boundary via:

$$A \# B = \lim_{\epsilon \downarrow 0} (A + \epsilon I) \# (B + \epsilon I).$$

But this extension is discontinuous [Bhatia-2007]: Let

$$A = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 20 & 6 \\ 6 & 2 \end{pmatrix}, X_n = \begin{pmatrix} 1 & 0 \\ 0 & 1/n \end{pmatrix} \rightarrow X = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$ 

Then, for $\Phi_n(A) = X_n^*AX_n$, $\Phi_n(A) \# \Phi_n(B) = \Phi_n(A \# B)$.

But sequential continuity is violated:

$$\lim_{n \uparrow \infty} \Phi_n(A) \# \Phi_n(B) = \lim_{n \uparrow \infty} \Phi_n(A \# B) = \Phi(A \# B) = \begin{pmatrix} 8 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\left( \lim_{n \uparrow \infty} \Phi_n(A) \right) \# \left( \lim_{n \uparrow \infty} \Phi_n(B) \right) = \Phi(A) \# \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$
Handling ill-conditioned Gramians [1/2]

The Nesterov-Todd equilibration obviously makes assumptions about the invertibility of the Gramians.

Geometrically, $S = A \# B$, when $A, B \in S_{++}^n$, is well-known to be the Euclidean midpoint between $\log(A)$ and $\log(B)$ and the midpoint of the geodesic between $A$ and $B$ when $S_{++}^n$ is equipped with the left-invariant metric $g_X(S, T) = \langle X^{-1}S, X^{-1}T \rangle$.

One could extend the geometric mean to the boundary via:

$$A \# B = \lim_{\epsilon \downarrow 0} (A + \epsilon I) \# (B + \epsilon I).$$

But this extension is discontinuous [Bhatia-2007]: Let

$$A = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 20 & 6 \\ 6 & 2 \end{pmatrix}, \quad X_n = \begin{pmatrix} 1 & 0 \\ 0 & 1/n \end{pmatrix} \rightarrow X = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Then, for $\Phi_n(A) = X_n^* A X_n$, $\Phi_n(A) \# \Phi_n(B) = \Phi_n(A \# B)$.

But sequential continuity is violated:

$$\lim_{n \uparrow \infty} \Phi_n(A) \# \Phi_n(B) = \lim_{n \uparrow \infty} \Phi_n(A \# B) = \Phi(A \# B) = \begin{pmatrix} 8 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\left(\lim_{n \uparrow \infty} \Phi_n(A)\right) \# \left(\lim_{n \uparrow \infty} \Phi_n(B)\right) = \Phi(A) \# \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$
Handling ill-conditioned Gramians [1/2]

The Nesterov-Todd equilibration obviously makes assumptions about the invertibility of the Gramians.

Geometrically, $S = A \# B$, when $A, B \in S^+_n$, is well-known to be the Euclidean midpoint between $\log(A)$ and $\log(B)$ and the midpoint of the geodesic between $A$ and $B$ when $S^+_n$ is equipped with the left-invariant metric $g_X(S, T) = \langle X^{-1}S, X^{-1}T \rangle$.

One could extend the geometric mean to the boundary via:

$$A \# B = \lim_{\epsilon \downarrow 0} (A + \epsilon I) \# (B + \epsilon I).$$

But this extension is discontinuous [Bhatia-2007]: Let

$$A = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 20 & 6 \\ 6 & 2 \end{pmatrix}, X_n = \begin{pmatrix} 1 & 0 \\ 0 & 1/n \end{pmatrix} \to X = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

Then, for $\Phi_n(A) = X_n^*AX_n$, $\Phi_n(A) \# \Phi_n(B) = \Phi_n(A \# B)$.

But sequential continuity is violated:

$$\lim_{n \uparrow \infty} \Phi_n(A) \# \Phi_n(B) = \lim_{n \uparrow \infty} \Phi_n(A \# B) = \Phi(A \# B) = \begin{pmatrix} 8 & 0 \\ 0 & 0 \end{pmatrix},$$

$$\left( \lim_{n \uparrow \infty} \Phi_n(A) \right) \# \left( \lim_{n \uparrow \infty} \Phi_n(B) \right) = \Phi(A) \# \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$
Handling ill-conditioned Gramians [2/2]

We thus saw that the extension:

\[ A \# B = \lim_{\epsilon \downarrow 0} (A + \epsilon I) \# (B + \epsilon I) \]

can lead to singular geometric means (in addition to being discontinuous).

But if we only care about \textbf{backwards stability}, then there is no issue. One can compute \( S = \hat{\hat{X}}^* \hat{\hat{X}}^{-1} \# \hat{\hat{Y}}^* \hat{\hat{Y}} \), where \( \hat{\hat{Z}} = Z + \alpha \|Z\|_F \) for some \( \alpha \ll 1 \), equilibrate with \( S \), and perhaps repeat.

This extends the applicability from \( S_{++}^n \) to \( S_+^n \setminus \{0\} \).
Handling ill-conditioned Gramians [2/2]

We thus saw that the extension:

\[ A \# B = \lim_{\epsilon \downarrow 0} (A + \epsilon I) \# (B + \epsilon I) \]

can lead to singular geometric means (in addition to being discontinuous).

But if we only care about backwards stability, then there is no issue. One can compute

\[ S = \hat{X}^* \hat{X}^{-1} \# \hat{Y}^* \hat{Y}, \]

where \( \hat{Z} = Z + \alpha \|Z\|_F \) for some \( \alpha \ll 1 \), equilibrate with \( S \), and perhaps repeat.

This extends the applicability from \( S^n_{++} \) to \( S^n_+ \setminus \{0\} \).
Handling ill-conditioned Gramians [2/2]

We thus saw that the extension:

\[ A \# B = \lim_{\epsilon \downarrow 0} (A + \epsilon I) \# (B + \epsilon I) \]

can lead to singular geometric means (in addition to being discontinuous).

But if we only care about **backwards stability**, then there is no issue. One can compute \( S = \hat{X}^* \hat{X}^{-1} \# \hat{Y}^* \hat{Y} \), where \( \hat{Z} = Z + \alpha \|Z\|_F \) for some \( \alpha \ll 1 \), equilibrate with \( S \), and perhaps repeat.

This extends the applicability from \( S^n_{++} \) to \( S^n_+ \setminus \{0\} \).
Another toy example

Consider minimizing $\|A - XY^*\|_F^2 + \lambda(\|X\|_F^2 + \|Y\|_F^2)$, given $A = \text{randn}(200, 400)$, $\lambda = 0.1$, $X_0 = \text{randn}(200, 10)$, $Y_0 = [\text{randn}(400, 9), \text{zeros}(400, 1)]$. 

![Log-loss graph]

- Blue line: Unequilibrated
- Red line: Equilibrated
Jordan-algebraic interpretations

Recall our definition $P(S) : \text{Sym}(n, \mathbb{R}) \to \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$.

This is a special case of the quadratic representation of a Jordan algebra $V$, where $P(x) = 2L(x)^2 - L(x^2)$ and $L(x) : V \to V$ is left application of $x \in V$.

For $V = \text{Sym}(n, \mathbb{R})$ with Jordan product $A \circ B \equiv \frac{1}{2}(AB + BA)$, $L(A)B \equiv A \circ B$:

$$P(A)B = 2(A \circ (A \circ B)) - A^2 \circ B = ABA.$$

The 1-to-1 correspondence between symmetric cones and squares of Euclidean Jordan algebras \cite{Faraut/Koranyi-1998} is commonly exploited in Interior Point Methods (especially for Lorentz cones).

One can easily build on Prop’n 1 to show: given $A, B \in \text{int}(V^2)$, there is a unique $S \in \text{int}(V^2)$ such that $P(S)A = B$. The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of $P$.

---

\footnote{\cite{Faraut/Koranyi-1998} Analysis on Symmetric Cones.} \footnote{\cite{Faybusovich-1997} Euclidean Jordan Algebras and Interior-point Alg’s} \footnote{\cite{Lim-2000} Geometric means on symmetric cones}
**Jordan-algebraic interpretations**

Recall our definition $P(S) : \text{Sym}(n, \mathbb{R}) \to \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$.

This is a special case of the quadratic representation of a **Jordan algebra** $V$, where $P(x) = 2L(x)^2 - L(x^2)$ and $L(x) : V \to V$ is left application of $x \in V$.

For $V = \text{Sym}(n, \mathbb{R})$ with Jordan product $A \circ B \equiv \frac{1}{2}(AB + BA)$, $L(A)B \equiv A \circ B$:

\[ P(A)B = 2(A \circ (A \circ B)) - A^2 \circ B = ABA. \]

The 1-to-1 correspondence between symmetric cones and squares of Euclidean Jordan algebras [Faraut/Koranyi-1998] is commonly exploited in Interior Point Methods (especially for Lorentz cones).

One can easily build on Prop’n 1 to show: given $A, B \in \text{int}(V^2)$, there is a unique $S \in \text{int}(V^2)$ such that $P(S)A = B$. The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of $P$.

---


7 [Faybusovich-1997] Euclidean Jordan Algebras and Interior-point Alg’s

**Jordan-algebraic interpretations**

Recall our definition $P(S) : \text{Sym}(n, \mathbb{R}) \rightarrow \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$. This is a special case of the quadratic representation of a **Jordan algebra** $V$, where $P(x) = 2L(x)^2 - L(x^2)$ and $L(x) : V \rightarrow V$ is left application of $x \in V$.

For $V = \text{Sym}(n, \mathbb{R})$ with Jordan product $A \circ B \equiv \frac{1}{2}(AB + BA)$, $L(A)B \equiv A \circ B$:

$$P(A)B = 2(A \circ (A \circ B)) - A^2 \circ B = ABA.$$ 

The 1-to-1 correspondence between symmetric cones and squares of Euclidean Jordan algebras [Faraut/Koranyi-1998] is commonly exploited in Interior Point Methods (especially for Lorentz cones).

One can easily build on Prop'n 1 to show: given $A, B \in \text{int}(V^2)$, there is a unique $S \in \text{int}(V^2)$ such that $P(S)A = B$. The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of $P$.

---


7 [Faybusovich-1997] Euclidean Jordan Algebras and Interior-point Alg’s

8 [Lim-2000] Geometric means on symmetric cones
Jordan-algebraic interpretations

Recall our definition $P(S) : \text{Sym}(n, \mathbb{R}) \to \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$.

This is a special case of the quadratic representation of a Jordan algebra $V$, where $P(x) = 2L(x)^2 - L(x^2)$ and $L(x) : V \to V$ is left application of $x \in V$.\footnote{[Faraut/Koranyi-1998] Analysis on Symmetric Cones.}

For $V = \text{Sym}(n, \mathbb{R})$ with Jordan product $A \circ B \equiv \frac{1}{2}(AB + BA)$, $L(A)B \equiv A \circ B$:

$$P(A)B = 2(A \circ (A \circ B)) - A^2 \circ B = ABA.$$\footnote{[Faybusovich-1997] Euclidean Jordan Algebras and Interior-point Alg’s}

The 1-to-1 correspondence between symmetric cones and squares of Euclidean Jordan algebras [Faraut/Koranyi-1998] is commonly exploited in Interior Point Methods (especially for Lorentz cones).\footnote{[Lim-2000] Geometric means on symmetric cones}

One can easily build on Prop’n 1 to show: given $A, B \in \text{int}(V^2)$, there is a unique $S \in \text{int}(V^2)$ such that $P(S)A = B$.\footnote{[Lim-2000] Geometric means on symmetric cones} The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of $P$. 

\footnote{[Faraut/Koranyi-1998] Analysis on Symmetric Cones.}
Jordan-algebraic interpretations

Recall our definition $P(S) : \text{Sym}(n, \mathbb{R}) \rightarrow \text{Sym}(n, \mathbb{R})$ via $P(S)A = SAS$.

This is a special case of the quadratic representation of a Jordan algebra $V$, where $P(x) = 2L(x)^2 - L(x^2)$ and $L(x) : V \rightarrow V$ is left application of $x \in V$.\(^6\)

For $V = \text{Sym}(n, \mathbb{R})$ with Jordan product $A \circ B \equiv \frac{1}{2}(AB + BA)$, $L(A)B \equiv A \circ B$:

$$ P(A)B = 2(A \circ (A \circ B)) - A^2 \circ B = ABA. $$

The 1-to-1 correspondence between symmetric cones and squares of Euclidean Jordan algebras [Faraut/Koranyi-1998] is commonly exploited in Interior Point Methods (especially for Lorentz cones).\(^7\)

One can easily build on Prop’n 1 to show: given $A, B \in \text{int}(V^2)$, there is a unique $S \in \text{int}(V^2)$ such that $P(S)A = B$.\(^8\) The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of $P$.

---


\(^7\)[Faybusovich-1997] Euclidean Jordan Algebras and Interior-point Alg’s

\(^8\)[Lim-2000] Geometric means on symmetric cones
**Determinantal Point Processes**

**Definition 6.** A **marginal kernel matrix** is a (real or complex) Hermitian matrix whose eigenvalues live in $[0, 1]$.

**Definition 7.** A **(finite) Determinantal Point Process (DPP)** is a random variable $Y$ over the power set of $\mathcal{Y} = \{1, \ldots, k\} \subset \mathbb{N}$ generated by a $k \times k$ marginal kernel matrix $K$ via the rule

$$P_K[Y \subseteq Y] = \det(K_Y),$$

where $K_Y$ is the $|Y| \times |Y|$ submatrix of $K$ formed by restricting to the rows and columns in the index set $Y$.

**Definition 8.** A DPP is called **elementary** if the eigenvalues of its marginal kernel matrix are all either 0 or 1.
Definition 6. A marginal kernel matrix is a (real or complex) Hermitian matrix whose eigenvalues live in [0, 1].

Definition 7. A (finite) Determinantal Point Process (DPP) is a random variable $Y$ over the power set of $\mathcal{Y} = \{1, \ldots, k\} \subseteq \mathbb{N}$ generated by a $k \times k$ marginal kernel matrix $K$ via the rule

$$P_K[Y \subseteq Y] = \det(K_Y),$$

where $K_Y$ is the $|Y| \times |Y|$ submatrix of $K$ formed by restricting to the rows and columns in the index set $Y$.

Definition 8. A DPP is called elementary if the eigenvalues of its marginal kernel matrix are all either 0 or 1.
**Determinantal Point Processes**

**Definition 6.** A **marginal kernel matrix** is a (real or complex) Hermitian matrix whose eigenvalues live in $[0, 1]$.

**Definition 7.** A (finite) **Determinantal Point Process (DPP)** is a random variable $Y$ over the power set of $\mathcal{Y} = \{1, \ldots, k\} \subset \mathbb{N}$ generated by a $k \times k$ marginal kernel matrix $K$ via the rule

$$P_K[Y \subseteq \mathcal{Y}] = \det(K_Y),$$

where $K_Y$ is the $|Y| \times |Y|$ submatrix of $K$ formed by restricting to the rows and columns in the index set $\mathcal{Y}$.

**Definition 8.** A DPP is called **elementary** if the eigenvalues of its marginal kernel matrix are all either 0 or 1.
How to sample a DPP?

Traditional algorithms [Hough et al.-2006] used an eigendecomposition of the kernel matrix and transformed the eigenvalues their Bernoulli draw to reduce to an elementary DPP (which was then sampled with a quartic algorithm).\(^9\)

[Gillenwater-2014] reduced the factored elementary DPP sampling down to cubic complexity via what is equivalent to diagonally-pivoted Cholesky.\(^10\)

Recently, authors are noticing the connections to Cholesky factorization for MAP inference and directly sampling from the marginal kernel. \(^11\)

I will give a simple proof of a cubic Cholesky-like algorithm for directly sampling from a marginal kernel and provide a high-performance blocked equivalent.

---


\(^10\) [Gillenwater-2014] Approximate inference for determinantal point processes

How to sample a DPP?

Traditional algorithms [Hough et al.-2006] used an eigendecomposition of the kernel matrix and transformed the eigenvalues their Bernoulli draw to reduce to an elementary DPP (which was then sampled with a quartic algorithm).\(^9\)

[Gillenwater-2014] reduced the factored elementary DPP sampling down to cubic complexity via what is equivalent to diagonally-pivoted Cholesky.\(^10\)

Recently, authors are noticing the connections to Cholesky factorization for MAP inference and directly sampling from the marginal kernel. \(^11\)

I will give a simple proof of a cubic Cholesky-like algorithm for directly sampling from a marginal kernel and provide a high-performance blocked equivalent.


\(^10\)[Gillenwater-2014] Approximate inference for determinantal point processes

How to sample a DPP?

Traditional algorithms [Hough et al.-2006] used an eigendecomposition of the kernel matrix and transformed the eigenvalues their Bernoulli draw to reduce to an elementary DPP (which was then sampled with a quartic algorithm).  

[Gillenwater-2014] reduced the factored elementary DPP sampling down to cubic complexity via what is equivalent to diagonally-pivoted Cholesky.  

Recently, authors are noticing the connections to Cholesky factorization for MAP inference and directly sampling from the marginal kernel.  

I will give a simple proof of a cubic Cholesky-like algorithm for directly sampling from a marginal kernel and provide a high-performance blocked equivalent.

---

10[Gillenwater-2014] Approximate inference for determinantal point processes  
How to sample a DPP?

Traditional algorithms [Hough et al.-2006] used an eigendecomposition of the kernel matrix and transformed the eigenvalues their Bernoulli draw to reduce to an elementary DPP (which was then sampled with a quartic algorithm). 9

[Gillenwater-2014] reduced the factored elementary DPP sampling down to cubic complexity via what is equivalent to diagonally-pivoted Cholesky. 10

Recently, authors are noticing the connections to Cholesky factorization for MAP inference and directly sampling from the marginal kernel. 11

I will give a simple proof of a cubic Cholesky-like algorithm for directly sampling from a marginal kernel and provide a high-performance blocked equivalent.

---

10 [Gillenwater-2014] Approximate inference for determinantal point processes
Complementary DPPs

Lemma 9 (Hough et al-2006). Given any $Y \sim \text{DPP}(K)$, where $K$ has spectral decomposition $Q\Lambda Q^*$, sampling from $Y$ is equivalent to sampling from the random elementary DPP with kernel $P(Q_Z)$, where $P(U) \equiv UU^*$ and $Q_Z$ consists of the columns of $Q$ with indices from $Z \sim \text{DPP}(\Lambda)$.

Lemma 10. Given any $Y \sim \text{DPP}(K)$, $Y^c \sim \text{DPP}(I - K)$ (which we call the complementary DPP). Proof. The case where $K$ is elementary is proven in [Tao-2009] via showing that the squared determinants of the diagonal blocks of a $2x2$ partition of an orthonormal matrix are equal.\(^{12}\)

In the general case, if $K$ has spectral decomposition $Q\Lambda Q^*$, then $I - K$ has spectral decomposition $Q(I - \Lambda)Q^*$. And the probability of drawing $J$ from DPP($\Lambda$) is equal to that of drawing $J^c$ from DPP($I - \Lambda$).

The result for the elementary case then shows that, if $Z \sim \text{DPP}(Q_jQ_j^*)$, then $Z^c \sim \text{DPP}(I - Q_jQ_j^*) = \text{DPP}(Q_j^cQ_j^{*c})$. The general case then follows from Lemma 9. □

---

\(^{12}\)[Tao-2009]
terrytao.wordpress.com/2009/08/23/determinantal-processes/
Complementary DPPs

**Lemma 9 (Hough et al-2006).** Given any $Y \sim DPP(K)$, where $K$ has spectral decomposition $Q\Lambda Q^*$, sampling from $Y$ is equivalent to sampling from the random elementary DPP with kernel $P(Q_Z)$, where $P(U) \equiv UU^*$ and $Q_Z$ consists of the columns of $Q$ with indices from $Z \sim DPP(\Lambda)$.

**Lemma 10.** Given any $Y \sim DPP(K)$, $Y^c \sim DPP(I - K)$ (which we call the **complementary DPP**). 

**Proof.** The case where $K$ is elementary is proven in [Tao-2009] via showing that the squared determinants of the diagonal blocks of a 2x2 partition of an orthonormal matrix are equal.$^{12}$

In the general case, if $K$ has spectral decomposition $Q\Lambda Q^*$, then $I - K$ has spectral decomposition $Q(I - \Lambda)Q^*$. And the probability of drawing $J$ from $DPP(\Lambda)$ is equal to that of drawing $J^c$ from $DPP(I - \Lambda)$.

The result for the elementary case then shows that, if $Z \sim DPP(Q_JQ_J^*)$, then $Z^c \sim DPP(I - Q_JQ_J^*) = DPP(Q_{J^c}Q_{J^c}^*)$. The general case then follows from Lemma 9. □

---

$^{12}$[Tao-2009]
terrytao.wordpress.com/2009/08/23/determinantal-processes/
Lemma 9 (Hough et al-2006). Given any $Y \sim \text{DPP}(K)$, where $K$ has spectral decomposition $Q\Lambda Q^*$, sampling from $Y$ is equivalent to sampling from the random elementary DPP with kernel $P(Q_z)$, where $P(U) \equiv UU^*$ and $Q_z$ consists of the columns of $Q$ with indices from $Z \sim \text{DPP}(\Lambda)$.

Lemma 10. Given any $Y \sim \text{DPP}(K)$, $Y^c \sim \text{DPP}(I - K)$ (which we call the complementary DPP). Proof. The case where $K$ is elementary is proven in [Tao-2009] via showing that the squared determinants of the diagonal blocks of a 2x2 partition of an orthonormal matrix are equal.\footnote{Tao-2009}

In the general case, if $K$ has spectral decomposition $Q\Lambda Q^*$, then $I - K$ has spectral decomposition $Q(I - \Lambda)Q^*$. And the probability of drawing $J$ from DPP$(\Lambda)$ is equal to that of drawing $J^c$ from DPP$(I - \Lambda)$.

The result for the elementary case then shows that, if $Z \sim \text{DPP}(Q_J Q_j^*)$, then $Z^c \sim \text{DPP}(I - Q_J Q_j^*) = \text{DPP}(Q_{J^c} Q_{j^\ast_c})$. The general case then follows from Lemma 9. □

\footnote{Tao-2009} terrytao.wordpress.com/2009/08/23/determinantal-processes/
Complementary DPPs

Lemma 9 (Hough et al-2006). Given any \( Y \sim \text{DPP}(K) \), where \( K \) has spectral decomposition \( Q \Lambda Q^* \), sampling from \( Y \) is equivalent to sampling from the random elementary DPP with kernel \( P(Q_Z) \), where \( P(U) \equiv UU^* \) and \( Q_Z \) consists of the columns of \( Q \) with indices from \( Z \sim \text{DPP}(\Lambda) \).

Lemma 10. Given any \( Y \sim \text{DPP}(K) \), \( Y^c \sim \text{DPP}(I - K) \) (which we call the complementary DPP). Proof. The case where \( K \) is elementary is proven in [Tao-2009] via showing that the squared determinants of the diagonal blocks of a 2x2 partition of an orthonormal matrix are equal.\(^{12}\)

In the general case, if \( K \) has spectral decomposition \( Q \Lambda Q^* \), then \( I - K \) has spectral decomposition \( Q(I - \Lambda)Q^* \). And the probability of drawing \( J \) from \( \text{DPP}(\Lambda) \) is equal to that of drawing \( J^c \) from \( \text{DPP}(I - \Lambda) \).

The result for the elementary case then shows that, if \( Z \sim \text{DPP}(Q_J Q_J^*) \), then \( Z^c \sim \text{DPP}(I - Q_J Q_J^*) = \text{DPP}(Q_{J^c} Q_{J^c}^*) \). The general case then follows from Lemma 9.

\(^{12}\) [Tao-2009]
terrytao.wordpress.com/2009/08/23/determinantal-processes/
Complementary DPPs

**Lemma 9 (Hough et al-2006).** Given any \( Y \sim \text{DPP}(K) \), where \( K \) has spectral decomposition \( Q\Lambda Q^* \), sampling from \( Y \) is equivalent to sampling from the random elementary DPP with kernel \( P(Q_Z) \), where \( P(U) \equiv UU^* \) and \( Q_Z \) consists of the columns of \( Q \) with indices from \( Z \sim \text{DPP}(\Lambda) \).

**Lemma 10.** Given any \( Y \sim \text{DPP}(K) \), \( Y^c \sim \text{DPP}(I - K) \) (which we call the complementary DPP). **Proof.** The case where \( K \) is elementary is proven in [Tao-2009] via showing that the squared determinants of the diagonal blocks of a 2x2 partition of an orthonormal matrix are equal.\(^{12}\)

In the general case, if \( K \) has spectral decomposition \( Q\Lambda Q^* \), then \( I - K \) has spectral decomposition \( Q(I - \Lambda)Q^* \). And the probability of drawing \( J \) from \( \text{DPP}(\Lambda) \) is equal to that of drawing \( J^c \) from \( \text{DPP}(I - \Lambda) \).

The result for the elementary case then shows that, if \( Z \sim \text{DPP}(Q_J Q^*_J) \), then \( Z^c \sim \text{DPP}(I - Q_J Q^*_J) = \text{DPP}(Q_{J^c} Q^*_{J^c}) \). The general case then follows from Lemma 9. \( \square \)

\(^{12}\) [Tao-2009]
terrytao.wordpress.com/2009/08/23/determinantal-processes/
Proposition 2. Given disjoint subsets \( A, B \subseteq \mathcal{Y} \),

\[
P[B \subseteq \mathcal{Y} | A \subseteq \mathcal{Y}] = \text{det}(K_B - K_{B,A}K_A^{-1}K_{A,B}),
\]

\[
P[B \subseteq \mathcal{Y} | A \subseteq \mathcal{Y}^c] = \text{det}(K_B + K_{B,A}(I - K_A)^{-1}K_{A,B}).
\]

Proof. The first claim follows from

\[
\text{det}(K_{A \cup B}) = \text{det}(K_A)\text{det}(K_B - K_{B,A}K_A^{-1}K_{A,B})
\]

and

\[
P[B \subseteq \mathcal{Y} | A \subseteq \mathcal{Y}] = \frac{\text{det}(K_{A \cup B})}{\text{det}(K_A)}.
\]

The second claim follows from applying the first result to the complementary DPP to find

\[
P[B \subseteq \mathcal{Y}^c | A \subseteq \mathcal{Y}^c] = \text{det}((I - K)_B - K_{B,A}(I - K)_A^{-1}K_{A,B}).
\]

Taking the complement of said Schur complement shows the second result. \( \square \)
Conditioning and Schur complements

Proposition 2. Given disjoint subsets $A, B \subset \mathcal{Y}$,

$$P[B \subseteq \mathcal{Y} | A \subseteq \mathcal{Y}] = \det(K_B - K_{B,A}K_A^{-1}K_{A,B}),$$

$$P[B \subseteq \mathcal{Y} | A \subseteq \mathcal{Y}^c] = \det(K_B + K_{B,A}(I - K_A)^{-1}K_{A,B}).$$

Proof. The first claim follows from

$$\det(K_{A \cup B}) = \det(K_A)\det(K_B - K_{B,A}K_A^{-1}K_{A,B})$$

and

$$P[B \subseteq \mathcal{Y} | A \subseteq \mathcal{Y}] = \frac{\det(K_{A \cup B})}{\det(K_A)}.$$

The second claim follows from applying the first result to the complementary DPP to find

$$P[B \subseteq \mathcal{Y}^c | A \subseteq \mathcal{Y}^c] = \det((I - K)_B - K_{B,A}(I - K)_A^{-1}K_{A,B}).$$

Taking the complement of said Schur complement shows the second result. □
Proposition 2. Given disjoint subsets \( A, B \subseteq \mathcal{Y} \),

\[
P[B \subseteq \mathcal{Y} | A \subseteq \mathcal{Y}] = \det(K_B - K_{B,A}K_A^{-1}K_{A,B}),
\]

\[
P[B \subseteq \mathcal{Y} | A \subseteq \mathcal{Y}^c] = \det(K_B + K_{B,A}(I - K_A)^{-1}K_{A,B}).
\]

Proof. The first claim follows from

\[
\det(K_{A \cup B}) = \det(K_A)\det(K_B - K_{B,A}K_A^{-1}K_{A,B})
\]

and

\[
P[B \subseteq \mathcal{Y} | A \subseteq \mathcal{Y}] = \frac{\det(K_{A \cup B})}{\det(K_A)}.
\]

The second claim follows from applying the first result to the complementary DPP to find

\[
P[B \subseteq \mathcal{Y}^c | A \subseteq \mathcal{Y}^c] = \det((I - K)_B - K_{B,A}(I - K)_A^{-1}K_{A,B}).
\]

Taking the complement of said Schur complement shows the second result. \( \square \)
Sampling w/ mirror-image Cholesky

```python
samples = {}
for j=1:n
    J2 = [j+1:n]
    keep_index = Bernoulli(K(j,j))
    if keep_index
        scale = -1; samples.insert(j)
        K(j,j) = sqrt(K(j,j))
    else
        scale = +1
        K(j,j) = sqrt(1-K(j,j))
    K(J2,j) /= K(j,j)
    K(J2,J2) += scale*tril(K(J2,j)*K(J2,j)')
```

This is a small tweak of unblocked Cholesky factorization; the majority of the work is in Hermitian rank-1 updates. And the standard Cholesky optimizations apply (e.g., blocking and sparse-direct factorization)!
Sampling w/ mirror-image Cholesky

```plaintext
samples = {}
for j=1:n
    J2 = [j+1:n]
    keep_index = Bernoulli(K(j,j))
    if keep_index
        scale = -1; samples.insert(j)
        K(j,j) = sqrt(K(j,j))
    else
        scale = +1
        K(j,j) = sqrt(1-K(j,j))
        K(J2,j) /= K(j,j)
        K(J2,J2) += scale*tril(K(J2,j)*K(J2,j)')
```

This is a small tweak of unblocked Cholesky factorization; the majority of the work is in Hermitian rank-1 updates. And the standard Cholesky optimizations apply (e.g., blocking and sparse-direct factorization)!
Blocked mirror-image sampling

\[
samples = {}
J1_{beg} = 1
\textbf{while} \quad J1_{beg} \leq n
\quad J1_{end} = \min(n, J1_{beg} + \text{blocksize} - 1)
J1 = [J1_{beg}:J1_{end}]; \quad J2 = [J1_{end} + 1:n]
J1_{samples}, \quad K(J1,J1) = \text{sample}(K(J1,J1))
A21 = \text{zeros}(\text{len}(J2), \text{len}(J1_{samples}))
B21 = \text{zeros}(\text{len}(J2), \text{len}(J1) - \text{len}(J1_{samples}))
\text{num\_keep\_packed} = \text{num\_drop\_packed} = 0
\textbf{for} \quad k \quad \text{in} \quad J1
\quad K(J2,k) /= K(k,k)
\quad \textbf{if} \quad (k - J1_{beg} + 1) \quad \text{in} \quad J1_{samples}
\quad \quad A21(:,\text{num\_keep\_packed}++) = K(J2,k); \quad \text{scale} = -1
\quad \textbf{else}
\quad \quad B21(:,\text{num\_drop\_packed}++) = K(J2,k); \quad \text{scale} = +1
\quad J1R = [k + 1:J1_{end}]
\quad K(J2,J1R) += \text{scale} * K(J2,k) * K(J1R,k)'
\quad K(J2,J2) += \text{tril}(B21*B21' - A21*A21')
\quad J1_{beg} = J1_{end} + 1
Dense single-core “Cholesky” sampling

HPC dense Cholesky implementations can be trivially modified.

Maximum Likelihood inference and elementary DPP sampling are similar but involve diagonal pivoting; the former uses the largest diagonal and the latter samples from the PDF implied by the diagonal. One can modify a blocked dense diagonally-pivoted Cholesky.

Sparse-direct Cholesky can be adapted for sampling a marginal kernel, but arbitrary pivoting can destroy its advantages for MAP and elementary DPPs.
HPC dense Cholesky implementations can be trivially modified.

Maximum Likelihood inference and elementary DPP sampling are similar but involve diagonal pivoting; the former uses the largest diagonal and the latter samples from the PDF implied by the diagonal. One can modify a blocked dense diagonally-pivoted Cholesky.

Sparse-direct Cholesky can be adapted for sampling a marginal kernel, but arbitrary pivoting can destroy its advantages for MAP and elementary DPPs.
Dense single-core “Cholesky” sampling

HPC dense Cholesky implementations can be trivially modified.

Maximum Likelihood inference and elementary DPP sampling are similar but involve diagonal pivoting; the former uses the largest diagonal and the latter samples from the PDF implied by the diagonal. One can modify a blocked dense diagonally-pivoted Cholesky.

Sparse-direct Cholesky can be adapted for sampling a marginal kernel, but arbitrary pivoting can destroy its advantages for MAP and elementary DPPs.
Dense single-core “Cholesky” sampling

HPC dense Cholesky implementations can be trivially modified.

Maximum Likelihood inference and elementary DPP sampling are similar but involve diagonal pivoting; the former uses the largest diagonal and the latter samples from the PDF implied by the diagonal. One can modify a blocked dense diagonally-pivoted Cholesky.

Sparse-direct Cholesky can be adapted for sampling a marginal kernel, but arbitrary pivoting can destroy its advantages for MAP and elementary DPPs.
Acknowledgements/Questions/Comments

Acknowledgements:

• Rasmus Larsen and John Anderson:
  For introducing me to the WALS problem.

• Steffen Rendle:
  For noticing that the Gramians were equal.

• Matt Knepley and Sameer Agarwal:
  For pointing out the gauge transformation analogy.

• Alex Kulesza and Jenny Gillenwater:
  For answering DPP sampling questions.

Questions/comments?