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:

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

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

[2] Cf. [Srebro/Jaakkola-2003] Weighted low-rank approximations

[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).[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

[2] Cf. [Srebro/Jaakkola-2003] Weighted low-rank approximations

[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).[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

[2] Cf. [Srebro/Jaakkola-2003] Weighted low-rank approximations

[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).[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

[2] Cf. [Srebro/Jaakkola-2003] Weighted low-rank approximations

[3] `http://www.tensorflow.org/api_docs/python/tf/contrib/factorization/WALSModel`

Why the Gramians are equivalent [1/3]

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

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

Proposition 1. For any $A, B \in S^n_{++}$, there is a unique $S \in S^n_{++}$ such that $P(S)A = B$.[4]

Proof. For existence, put $S = A^{-1} \sharp B$.
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 $S = A^{-1} \sharp B$.[5]

---

[4] [Anderson/Trapp-1980] Operator means and electrical networks, Cf. [Bhatia-2007] Positive Definite Matrices

[5] [Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones

Why the Gramians are equivalent [1/3]

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

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

Proposition 1. For any $A, B \in S^n_{++}$, there is a unique $S \in S^n_{++}$ such that $P(S)A = B$.[4]

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

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 $S = A^{-1} \sharp B$.[5]

[4] [Anderson/Trapp-1980] Operator means and electrical networks, Cf. [Bhatia-2007] Positive Definite Matrices

[5] [Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones

Why the Gramians are equivalent [1/3]

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

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

Proposition 1. For any $A, B \in S^n_{++}$, there is a unique $S \in S^n_{++}$ such that $P(S)A = B$.[4]

Proof. For existence, put $S = A^{-1} \sharp B$.
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 $S = A^{-1} \sharp B$.[5]

[4] [Anderson/Trapp-1980] Operator means and electrical networks, Cf. [Bhatia-2007] Positive Definite Matrices

[5] [Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones

Why the Gramians are equivalent [1/3]

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

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

Proposition 1. For any $A, B \in S^n_{++}$, there is a unique $S \in S^n_{++}$ such that $P(S)A = B$.[4]

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

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 $S = A^{-1} \sharp B$.[5]

---

[4] [Anderson/Trapp-1980] Operator means and electrical networks, Cf. [Bhatia-2007] Positive Definite Matrices

[5] [Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones

Why the Gramians are equivalent [1/3]

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

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

Proposition 1. For any $A, B \in S^n_{++}$, there is a unique $S \in S^n_{++}$ such that $P(S)A = B$.[4]
Proof. For existence, put $S = A^{-1} \sharp B$.
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 $S = A^{-1} \sharp B$.[5]

[4] [Anderson/Trapp-1980] Operator means and electrical networks, Cf. [Bhatia-2007] Positive Definite Matrices

[5] [Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones

Why the Gramians are equivalent [1/3]

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

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

Proposition 1. For any $A, B \in S^n_{++}$, there is a unique $S \in S^n_{++}$ such that $P(S)A = B$.[4]

Proof. For existence, put $S = A^{-1} \sharp B$.
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 $S = A^{-1} \sharp B$.[5]

---

[4] [Anderson/Trapp-1980] Operator means and electrical networks, Cf. [Bhatia-2007] Positive Definite Matrices

[5] [Nesterov/Todd-1998] Primal-Dual Interior Point Methods for self-scaled cones

Why the Gramians are equivalent [2/3]

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 $S = ((X^*X)^{-1} \sharp (Y^*Y))^{1/2}$ 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 $dg_T : (T_T S^n_{++} \cong \mathrm{Sym}(n, \mathbb{R})) \to (T_{g(T)}\mathbb{R} \cong \mathbb{R})$ 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 [2/3]

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 $S = ((X^*X)^{-1} \sharp (Y^*Y))^{1/2}$ 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 $dg_T : (T_T S^n_{++} \cong \mathrm{Sym}(n,\mathbb{R})) \to (T_{g(T)}\mathbb{R} \cong \mathbb{R})$ 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 [2/3]

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 $S = ((X^*X)^{-1} \sharp (Y^*Y))^{1/2}$ 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 $dg_T : (T_T S^n_{++} \cong \mathrm{Sym}(n, \mathbb{R})) \to (T_{g(T)}\mathbb{R} \cong \mathbb{R})$ 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 $X^*X - S^{-2}Y^*YS^{-2} = 0$. □

Why the Gramians are equivalent [2/3]

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 $S = ((X^*X)^{-1} \sharp (Y^*Y))^{1/2}$ 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 $dg_T : (T_T S^n_{++} \cong \mathrm{Sym}(n,\mathbb{R})) \to (T_{g(T)}\mathbb{R} \cong \mathbb{R})$ 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 $S = (X^*X)^{-1} \sharp (Y^*Y)$, minimizes the regularization while preserving the input to ℓ.

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

$$\begin{aligned} \|X\|_F^2 + \|Y\|_F^2 &= \min_{Z\in GL(n,\mathbb{R})} \{\|XZ\|_F^2 + \|YZ^{-*}\|_F^2\} \\ &= \min_{S\in S^n_{++}} \{\|XS\|_F^2 + \|YS^{-1}\|_F^2\}, \end{aligned}$$

where we exploited the polar decomposition $Z = SQ$, Q unitary. The result then follows from our lemma. □

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 $S = (X^*X)^{-1} \sharp (Y^*Y)$, minimizes the regularization while preserving the input to ℓ.

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

$$\begin{aligned}\|X\|_F^2 + \|Y\|_F^2 &= \min_{Z\in GL(n,\mathbb{R})} \{\|XZ\|_F^2 + \|YZ^{-*}\|_F^2\} \\ &= \min_{S\in S^n_{++}} \{\|XS\|_F^2 + \|YS^{-1}\|_F^2\},\end{aligned}$$

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 $S = (X^*X)^{-1} \sharp (Y^*Y)$, minimizes the regularization while preserving the input to ℓ.

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

$$\begin{aligned}
\|X\|_F^2 + \|Y\|_F^2 &= \min_{Z\in GL(n,\mathbb{R})} \{\|XZ\|_F^2 + \|YZ^{-*}\|_F^2\} \\
&= \min_{S\in S_{++}^n} \{\|XS\|_F^2 + \|YS^{-1}\|_F^2\},
\end{aligned}$$

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 $S = (X^*X)^{-1} \sharp (Y^*Y)$, minimizes the regularization while preserving the input to ℓ.

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

$$\begin{aligned}\|X\|_F^2 + \|Y\|_F^2 &= \min_{Z\in GL(n,\mathbb{R})}\{\|XZ\|_F^2 + \|YZ^{-*}\|_F^2\}\\ &= \min_{S\in S^n_{++}}\{\|XS\|_F^2 + \|YS^{-1}\|_F^2\},\end{aligned}$$

where we exploited the polar decomposition $Z = SQ$, Q unitary. The result then follows from our lemma. □

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

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

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

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

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

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

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

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

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

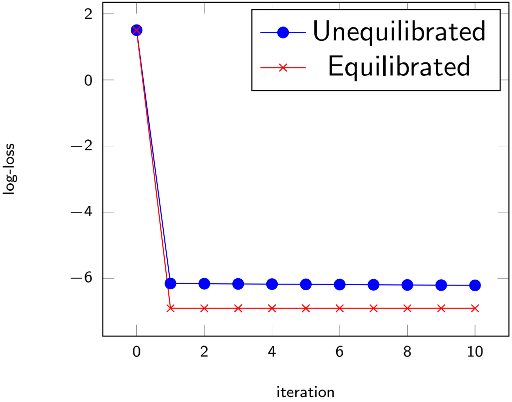
Equilibration is essentially free and keeps the regularization minimized (with the constraint of preserving the loss function input).

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, $S = A \sharp B$, 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:

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

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 $\Phi_n(A) = X_n^* A X_n$, $\Phi_n(A) \sharp \Phi_n(B) = \Phi_n(A \sharp B)$.
But sequential continuity is violated:

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

$$\left(\lim_{n \uparrow \infty} \Phi_n(A) \right) \sharp \left(\lim_{n \uparrow \infty} \Phi_n(B) \right) = \Phi(A) \sharp \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$

Handling ill-conditioned Gramians [1/2]

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

Geometrically, $S = A \sharp B$, 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:

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

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 $\Phi_n(A) = X_n^* A X_n$, $\Phi_n(A) \sharp \Phi_n(B) = \Phi_n(A \sharp B)$.
But sequential continuity is violated:

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

$$\left(\lim_{n \uparrow \infty} \Phi_n(A) \right) \sharp \left(\lim_{n \uparrow \infty} \Phi_n(B) \right) = \Phi(A) \sharp \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$

Handling ill-conditioned Gramians [1/2]

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

Geometrically, $S = A \sharp B$, 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:

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

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 $\Phi_n(A) = X_n^* A X_n$, $\Phi_n(A) \sharp \Phi_n(B) = \Phi_n(A \sharp B)$.
But sequential continuity is violated:

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

$$\left(\lim_{n \uparrow \infty} \Phi_n(A) \right) \sharp \left(\lim_{n \uparrow \infty} \Phi_n(B) \right) = \Phi(A) \sharp \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$

Handling ill-conditioned Gramians [1/2]

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

Geometrically, $S = A \sharp B$, 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:

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

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 $\Phi_n(A) = X_n^* A X_n$, $\Phi_n(A) \sharp \Phi_n(B) = \Phi_n(A \sharp B)$.
But sequential continuity is violated:

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

$$\left(\lim_{n \uparrow \infty} \Phi_n(A) \right) \sharp \left(\lim_{n \uparrow \infty} \Phi_n(B) \right) = \Phi(A) \sharp \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$

Handling ill-conditioned Gramians [1/2]

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

Geometrically, $S = A \sharp B$, 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:

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

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 $\Phi_n(A) = X_n^* A X_n$, $\Phi_n(A) \sharp \Phi_n(B) = \Phi_n(A \sharp B)$.
But sequential continuity is violated:

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

$$\left(\lim_{n \uparrow \infty} \Phi_n(A) \right) \sharp \left(\lim_{n \uparrow \infty} \Phi_n(B) \right) = \Phi(A) \sharp \Phi(B) = \begin{pmatrix} \sqrt{80} & 0 \\ 0 & 0 \end{pmatrix}.$$

Handling ill-conditioned Gramians [2/2]

We thus saw that the extension:

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

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 $S = \widehat{X^*X}^{-1} \sharp \widehat{Y^*Y}$, 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:

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

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 $S = \widehat{X^* X}^{-1} \sharp \widehat{Y^* Y}$, 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:

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

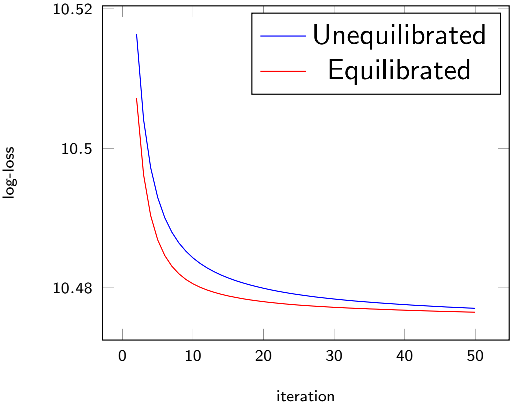
can lead to singular geometric means (in addition to being discontinuous).

But if we only care about **backwards stability**, then there is no issue. One can compute $S = \widehat{X^* X}^{-1} \sharp \widehat{Y^* Y}$, 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 $A = \mathrm{randn}(200, 400)$, $\lambda = 0.1$, $X_0 = \mathrm{randn}(200, 10)$, $Y_0 = [\mathrm{randn}(400, 9), \mathrm{zeros}(400, 1)]$.



Jordan-algebraic interpretations

Recall our definition $P(S) : \mathrm{Sym}(n, \mathbb{R}) \to \mathrm{Sym}(n, \mathbb{R})$ 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 $x \in V$.[6]

For $V = \mathrm{Sym}(n, \mathbb{R})$ 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 $A, B \in \mathrm{int}(V^2)$, there is a unique $S \in \mathrm{int}(V^2)$ such that $P(S)A = B$.[8] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

[6] [Faraut/Koranyi-1998] Analysis on Symmetric Cones.
[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 $P(S) : \mathrm{Sym}(n, \mathbb{R}) \to \mathrm{Sym}(n, \mathbb{R})$ 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 $x \in V$.[6]

For $V = \mathrm{Sym}(n, \mathbb{R})$ 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 $A, B \in \mathrm{int}(V^2)$, there is a unique $S \in \mathrm{int}(V^2)$ such that $P(S)A = B$.[8] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

[6] [Faraut/Koranyi-1998] Analysis on Symmetric Cones.

[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 $P(S) : \mathrm{Sym}(n, \mathbb{R}) \to \mathrm{Sym}(n, \mathbb{R})$ 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 $x \in V$.[6]

For $V = \mathrm{Sym}(n, \mathbb{R})$ 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 $A, B \in \mathrm{int}(V^2)$, there is a unique $S \in \mathrm{int}(V^2)$ such that $P(S)A = B$.[8] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

---

[6] [Faraut/Koranyi-1998] Analysis on Symmetric Cones.
[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 $P(S) : \mathrm{Sym}(n, \mathbb{R}) \to \mathrm{Sym}(n, \mathbb{R})$ 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 $x \in V$.[6]

For $V = \mathrm{Sym}(n, \mathbb{R})$ 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 $A, B \in \mathrm{int}(V^2)$, there is a unique $S \in \mathrm{int}(V^2)$ such that $P(S)A = B$.[8] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

[6] [Faraut/Koranyi-1998] Analysis on Symmetric Cones.

[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 $P(S) : \mathrm{Sym}(n, \mathbb{R}) \to \mathrm{Sym}(n, \mathbb{R})$ 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 $x \in V$.[6]

For $V = \mathrm{Sym}(n, \mathbb{R})$ 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 $A, B \in \mathrm{int}(V^2)$, there is a unique $S \in \mathrm{int}(V^2)$ such that $P(S)A = B$.[8] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

[6] [Faraut/Koranyi-1998] Analysis on Symmetric Cones.
[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 $\mathbf{Y}$ over the power set of $\mathcal{Y} = \{1, ..., k\} \subset \mathbb{N}$ generated by a $k \times k$ marginal kernel matrix K via the rule

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

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 $\mathbf{Y}$ over the power set of $\mathcal{Y} = \{1, ..., k\} \subset \mathbb{N}$ generated by a $k \times k$ marginal kernel matrix K via the rule

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

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 $\mathbf{Y}$ over the power set of $\mathcal{Y} = \{1, ..., k\} \subset \mathbb{N}$ generated by a $k \times k$ marginal kernel matrix K via the rule

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

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.

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.

[9] [Hough et al.-2006] Determinantal point processes and independence, Cf. [Kulesza/Taskar-2012] Determinantal point processes for machine learning.

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

[11] [Chen et al.-2017] Fast Greedy MAP inference for Determinantal Point Processes, [Launay et al.-2018] Exact sampling of determinantal point processes without eigendecomposition

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.

[9] [Hough et al.-2006] Determinantal point processes and independence, Cf. [Kulesza/Taskar-2012] Determinantal point processes for machine learning.

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

[11] [Chen et al.-2017] Fast Greedy MAP inference for Determinantal Point Processes, [Launay et al.-2018] Exact sampling of determinantal point processes without eigendecomposition

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.

---

[9] [Hough et al.-2006] Determinantal point processes and independence, Cf. [Kulesza/Taskar-2012] Determinantal point processes for machine learning.

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

[11] [Chen et al.-2017] Fast Greedy MAP inference for Determinantal Point Processes, [Launay et al.-2018] Exact sampling of determinantal point processes without eigendecomposition

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.

---

[9] [Hough et al.-2006] Determinantal point processes and independence, Cf. [Kulesza/Taskar-2012] Determinantal point processes for machine learning.

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

[11] [Chen et al.-2017] Fast Greedy MAP inference for Determinantal Point Processes, [Launay et al.-2018] Exact sampling of determinantal point processes without eigendecomposition

Complementary DPPs

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

Lemma 10. Given any $\mathbf{Y} \sim \mathrm{DPP}(K)$, $\mathbf{Y}^c \sim \mathrm{DPP}(I - K)$ (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 $\mathrm{DPP}(\Lambda)$ is equal to that of drawing J^c from $\mathrm{DPP}(I - \Lambda)$.

The result for the elementary case then shows that, if $\mathbf{Z} \sim \mathrm{DPP}(Q_J Q_J^*)$, then $\mathbf{Z}^c \sim \mathrm{DPP}(I - Q_J Q_J^*) = \mathrm{DPP}(Q_{J^c} Q_{J^c}^*)$. 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 $\mathbf{Y} \sim \mathrm{DPP}(K)$, where K has spectral decomposition $Q\Lambda Q^*$, sampling from $\mathbf{Y}$ is equivalent to sampling from the random elementary DPP with kernel $P(Q_{\mathbf{Z}})$, where $P(U) \equiv UU^*$ and $Q_{\mathbf{Z}}$ consists of the columns of Q with indices from $\mathbf{Z} \sim \mathrm{DPP}(\Lambda)$.

Lemma 10. Given any $\mathbf{Y} \sim \mathrm{DPP}(K)$, $\mathbf{Y}^c \sim \mathrm{DPP}(I-K)$ (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 $\mathrm{DPP}(\Lambda)$ is equal to that of drawing J^c from $\mathrm{DPP}(I-\Lambda)$.

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

[12] [Tao-2009] terrytao.wordpress.com/2009/08/23/determinantal-processes/

Complementary DPPs

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

Lemma 10. Given any $\mathbf{Y} \sim \text{DPP}(K)$, $\mathbf{Y}^c \sim \text{DPP}(I - K)$ (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 $\mathbf{Z} \sim \text{DPP}(Q_J Q_J^*)$, then $\mathbf{Z}^c \sim \text{DPP}(I - Q_J Q_J^*) = \text{DPP}(Q_{J^c} Q_{J^c}^*)$. The general case then follows from Lemma 9. $\square$

[12] [Tao-2009]
`terrytao.wordpress.com/2009/08/23/determinantal-processes/`

Complementary DPPs

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

Lemma 10. Given any $\mathbf{Y} \sim \mathrm{DPP}(K)$, $\mathbf{Y}^c \sim \mathrm{DPP}(I-K)$ (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 $\mathrm{DPP}(\Lambda)$ is equal to that of drawing J^c from $\mathrm{DPP}(I-\Lambda)$.

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

[12] [Tao-2009]
terrytao.wordpress.com/2009/08/23/determinantal-processes/

Complementary DPPs

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

Lemma 10. Given any $\mathbf{Y} \sim \mathrm{DPP}(K)$, $\mathbf{Y}^c \sim \mathrm{DPP}(I - K)$ (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 $\mathrm{DPP}(\Lambda)$ is equal to that of drawing J^c from $\mathrm{DPP}(I - \Lambda)$.

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

[12] [Tao-2009] terrytao.wordpress.com/2009/08/23/determinantal-processes/

Conditioning and Schur complements

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

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

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

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

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

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

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

Taking the complement of said Schur complement shows the second result. □

Conditioning and Schur complements

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

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

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

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

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

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

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

Taking the complement of said Schur complement shows the second result. □

Conditioning and Schur complements

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

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

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

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

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

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

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

Taking the complement of said Schur complement shows the second result. $\square$

Sampling w/ mirror-image Cholesky

```
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

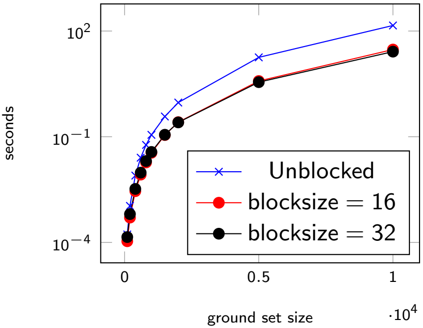
```
samples = {}
for j=1:n
  J2 = [j+1:n]
  keep_index = Bernoulli(K(j,j))
  if keep_index
    scale = -1; samples.insert(j)
    K(j,j) = sqrt(K(j,j))
  else
    scale = +1
    K(j,j) = sqrt(1-K(j,j))
  K(J2,j) /= K(j,j)
  K(J2,J2) += scale*tril(K(J2,j)*K(J2,j)')
```

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
while J1_beg <= n
  J1_end = min(n, J1_beg+blocksize-1)
  J1 = [J1_beg:J1_end]; J2 = [J1_end+1:n]
  J1_samples, K(J1,J1) = sample(K(J1,J1))
  A21 = zeros(len(J2), len(J1_samples))
  B21 = zeros(len(J2), len(J1)-len(J1_samples))
  num_keep_packed = num_drop_packed = 0
  for k in J1
    K(J2,k) /= K(k,k)
    if (k-J1_beg+1) in J1_samples
      A21(:,num_keep_packed++) = K(J2,k); scale = -1
    else
      B21(:,num_drop_packed++) = K(J2,k); scale = +1
    J1R = [k+1:J1_end]
    K(J2,J1R) += scale*K(J2,k)*K(J1R,k)'
  K(J2,J2) += tril(B21*B21' - A21*A21')
  J1_beg = J1_end + 1
```

Dense single-core "Cholesky" sampling

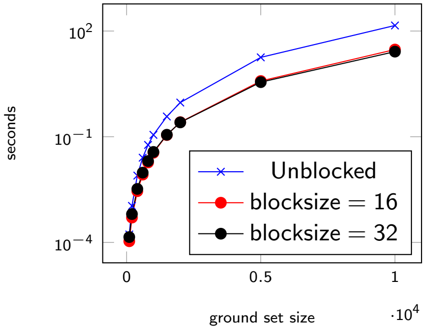


HPC dense Cholesky implementations can be trivially modified.

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.

Dense single-core "Cholesky" sampling

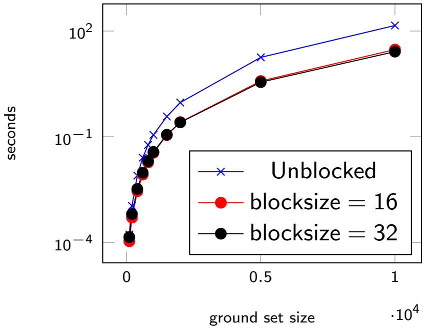


HPC dense Cholesky implementations can be trivially modified.

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

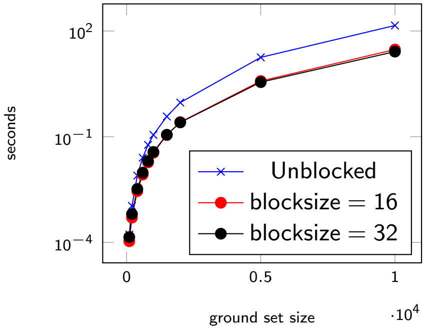


HPC dense Cholesky implementations can be trivially modified.

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.

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?