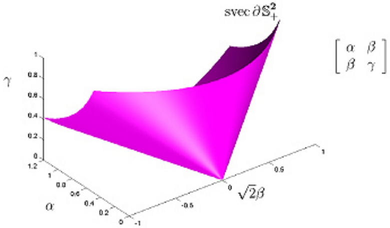


Equilibrating low-rank approximations with Gaussian priors

Jack Poulson (Hodge Star Scientific Computing)
Aussois, France, June 21, 2019

June 21, 2019

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 $T = S$. □

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 $T = S$. □

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 $T = S$. □

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 $X^*X - S^{-2}Y^*YS^{-2} = 0$. □

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 $X^*X - S^{-2}Y^*YS^{-2} = 0$. □

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 [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. □

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. □

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.

Brief SDP IPM intro

A **semidefinite program** can always be represented in the form:

$$\arg \inf_{x, S_0, \ldots, S_{N-1}} \{ c^T x : Ax = b,$$

$$G_k x + \mathrm{vec}(S_k) = \mathrm{vec}(H_k), S_k \succeq 0, \; k = 0, \ldots, N-1\},$$

where $S_k, H_k \in S^{n_k}$.

In the case where each $n_k = 1$, the above reduces to a **linear program**:

$$\arg \inf_{x,s} \{c^T x : Ax = b, Gx + s = h, s \geq 0\}.$$

In the case where $N = 1$, we have the SDP

$$\arg \inf_{x,S} \{c^T x : Ax = b, Gx + \mathrm{vec}(S) = \mathrm{vec}(H), S \succeq 0\}.$$

We will write $s := \mathrm{vec}(S)$ and $h := \mathrm{vec}(H)$ for brevity.

Brief SDP IPM intro

A **semidefinite program** can always be represented in the form:

$$\arg\inf_{x,S_0,\ldots,S_{N-1}} \{ c^T x : Ax = b,$$
$$G_k x + \mathrm{vec}(S_k) = \mathrm{vec}(H_k), S_k \succeq 0,\ k = 0, \ldots, N-1\},$$

where $S_k, H_k \in S^{n_k}$.

In the case where each $n_k = 1$, the above reduces to a **linear program**:

$$\arg\inf_{x,s}\{c^T x : Ax = b, Gx + s = h, s \geq 0\}.$$

In the case where $N = 1$, we have the SDP

$$\arg\inf_{x,S}\{c^T x : Ax = b, Gx + \mathrm{vec}(S) = \mathrm{vec}(H), S \succeq 0\}.$$

We will write $s := \mathrm{vec}(S)$ and $h := \mathrm{vec}(H)$ for brevity.

Brief SDP IPM intro

A **semidefinite program** can always be represented in the form:

$$\arg \inf_{x, S_0, \ldots, S_{N-1}} \quad \{ c^T x : Ax = b, \\ G_k x + \mathrm{vec}(S_k) = \mathrm{vec}(H_k), S_k \succeq 0, \ k = 0, \ldots, N-1 \},$$

where $S_k, H_k \in S^{n_k}$.

In the case where each $n_k = 1$, the above reduces to a **linear program**:

$$\arg \inf_{x,s} \{ c^T x : Ax = b, Gx + s = h, s \geq 0 \}.$$

In the case where $N = 1$, we have the SDP

$$\arg \inf_{x,S} \{ c^T x : Ax = b, Gx + \mathrm{vec}(S) = \mathrm{vec}(H), S \succeq 0 \}.$$

We will write $s := \mathrm{vec}(S)$ and $h := \mathrm{vec}(H)$ for brevity.

Brief SDP IPM intro

In the $N = 1$ case, the SDP

$$\arg\inf_{x,S}\{c^T x : Ax = b, Gx + \mathrm{vec}(S) = \mathrm{vec}(H), S \succeq 0\}.$$

has a Lagrangian

$$\mathcal{L}(x, S; y, Z) = c^T x + y^T (Ax - b) + z^T (Gx + s - h),$$

under the constraint $S \succeq 0$, where we put $z := \mathrm{vec}(Z)$.

Introducing the **barrier function** $\Phi(S) = -\ln(\det(S))$, and a **barrier parameter** $\mu > 0$, we have the unconstrained Lagrangian

$$\mathcal{L}_\mu(x, S; y, Z) = c^T x + y^T (Ax - b) + z^T (Gx + s - h) + \mu\Phi(S).$$

Brief SDP IPM intro

In the $N = 1$ case, the SDP

$$\arg\inf_{x,S}\{c^T x : Ax = b, Gx + \text{vec}(S) = \text{vec}(H), S \succeq 0\}.$$

has a Lagrangian

$$\mathcal{L}(x, S; y, Z) = c^T x + y^T(Ax - b) + z^T(Gx + s - h),$$

under the constraint $S \succeq 0$, where we put $z := \text{vec}(Z)$.

Introducing the **barrier function** $\Phi(S) = -\ln(\det(S))$, and a **barrier parameter** $\mu > 0$, we have the unconstrained Lagrangian

$$\mathcal{L}_\mu(x, S; y, Z) = c^T x + y^T(Ax - b) + z^T(Gx + s - h) + \mu\Phi(S).$$

Brief SDP IPM intro

$$\mathcal{L}_\mu(x, S; y, Z) = c^T x + y^T(Ax - b) + z^T(Gx + s - h) + \mu\Phi(S).$$

The critical points of this unconstrained Lagrangian are:

$$\begin{aligned} d_x(\mathcal{L}_\mu) &= c + A^T y + G^T z = 0, \\ d_y(\mathcal{L}_\mu) &= Ax - b = 0, \\ d_Z(\mathcal{L}_\mu) &= Gx + s - h = 0, \\ d_S(\mathcal{L}_\mu) &= Z - \mu d_s(\log(\det(S))) = Z - \mu S^{-1} = 0, \end{aligned}$$

[6] so that the last equation implies the **complementarity condition**

$$SZ = ZS = \mu I.$$

All first-order optimality conditions are linear except for the complementarity condition.

[6] Consider the curve $\phi(t) = S + tA$...

Brief SDP IPM intro

$$\mathcal{L}_\mu(x, S; y, Z) = c^T x + y^T(Ax - b) + z^T(Gx + s - h) + \mu\Phi(S).$$

The critical points of this unconstrained Lagrangian are:

$$\begin{aligned} d_x(\mathcal{L}_\mu) &= c + A^T y + G^T z = 0, \\ d_y(\mathcal{L}_\mu) &= Ax - b = 0, \\ d_Z(\mathcal{L}_\mu) &= Gx + s - h = 0, \\ d_S(\mathcal{L}_\mu) &= Z - \mu d_s(\log(\det(S))) = Z - \mu S^{-1} = 0, \end{aligned}$$

[6] so that the last equation implies the **complementarity condition**

$$SZ = ZS = \mu I.$$

All first-order optimality conditions are linear except for the complementarity condition.

[6] Consider the curve $\phi(t) = S + tA$...

Brief SDP IPM intro

Consider an **inner automorphism** $S \mapsto XSX$, for some $X \in S^n$. Then the complementarity condition equals

$$(XSX)(X^{-1}ZX^{-1}) = XSZX^{-1} = X(\mu I)X^{-1} = \mu I,$$

so that we are free to choose any such automorphism $(S, Z) \mapsto (XSX, X^{-1}ZX^{-1})$ before linearizing.

As we proved, there is a unique $X \in S^n_{++}$ such that $XSX = X^{-1}ZX^{-1} = W \in S^n_{++}$, where W is the Nesterov-Todd scaling point of the primal-dual pair (S, Z).

Brief SDP IPM intro

Consider an **inner automorphism** $S \mapsto XSX$, for some $X \in S^n$. Then the complementarity condition equals

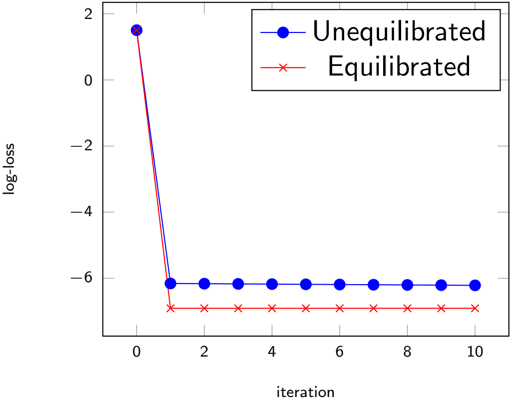
$$(XSX)(X^{-1}ZX^{-1}) = XSZX^{-1} = X(\mu I)X^{-1} = \mu I,$$

so that we are free to choose any such automorphism $(S, Z) \mapsto (XSX, X^{-1}ZX^{-1})$ before linearizing.

As we proved, there is a unique $X \in S^n_{++}$ such that $XSX = X^{-1}ZX^{-1} = W \in S^n_{++}$, where W is the Nesterov-Todd scaling point of the primal-dual pair (S, Z).

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

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

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 [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:

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

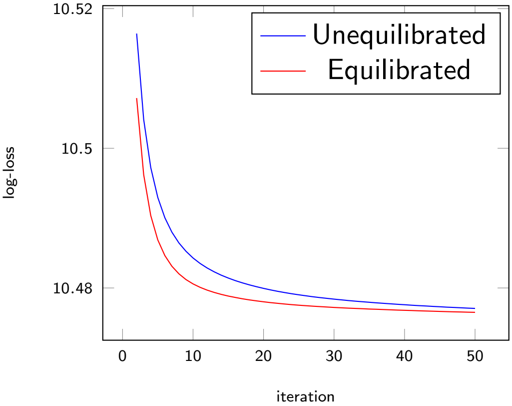
can lead to singular geometric means (in addition to being discontinuous).

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 $A = \text{randn}(200, 400)$, $\lambda = 0.1$, $X_0 = \text{randn}(200, 10)$, $Y_0 = [\text{randn}(400, 9), \text{zeros}(400, 1)]$.



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 $x \in V$.[7]

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).[8]

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 $P(S)A = B$.[9] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

[7] [Faraut/Koranyi-1998] Analysis on Symmetric Cones.
[8] [Faybusovich-1997] Euclidean Jordan Algebras and Interior-point Alg's
[9] [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 $x \in V$.[7]

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).[8]

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 $P(S)A = B$.[9] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

[7] [Faraut/Koranyi-1998] Analysis on Symmetric Cones.
[8] [Faybusovich-1997] Euclidean Jordan Algebras and Interior-point Alg's
[9] [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 $x \in V$.[7]

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).[8]

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 $P(S)A = B$.[9] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

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

[8] [Faybusovich-1997] Euclidean Jordan Algebras and Interior-point Alg's

[9] [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 $x \in V$.[7]

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).[8]

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 $P(S)A = B$.[9] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

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

[8] [Faybusovich-1997] Euclidean Jordan Algebras and Interior-point Alg's

[9] [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 $x \in V$.[7]

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).[8]

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 $P(S)A = B$.[9] The definitions of geometric means and Nesterov-Todd scaling points carry over through usage of P.

[7] [Faraut/Koranyi-1998] Analysis on Symmetric Cones.
[8] [Faybusovich-1997] Euclidean Jordan Algebras and Interior-point Alg's
[9] [Lim-2000] Geometric means on symmetric cones

Discussion

These slides are available at:
hodgestar.com/G2S3/

Questions/comments?
Chatroom at:

`https://gitter.im/hodge_star/G2S3`

Lab 3: Equilibrated, diverse recommendations

1. Insert transformation of Gramians to their Nesterov-Todd scaling point after each update and print objective function after each step.
2. Return the nearest 50 neighbors of our previous examples: "france", "music", "holiday", "summer", and "mountain".
3. Sample 10 terms from a DPP over the nearest 50 neighbors of each term via a marginal kernel of the form:

$$K_{i,j} = \gamma \begin{cases} \cos(\text{query}, \text{candidate}_i)^p, & i = j \\ \alpha \text{candidate}_i^T \text{candidate}_j, & i \neq j \end{cases}$$

   for various values of $\alpha \geq 0$ and $p > 0$ – checking for the matrix being positive-semidefinite then rescaling to have a preferred norm, reporting the most interesting combinations of parameters and results.