Matrix Algebra and Optimization for Statistics and Machine Learning

Yiyuan She

Department of Statistics, Florida State University

- Convexity and convex optimization

Convex functions

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathrm{dom} f$ is a convex set and

$$\theta f(x) + (1-\theta) f(y) - f(\theta x + (1-\theta) y) \geq 0$$

  holds for any $x, y \in \mathrm{dom} f$ and any $\theta \in [0,1]$.
  - Equivalently, $f(x+tv)$ as a function of $t \in \mathbb{R}$ is convex
  - The RHS defines a useful operator $\mathbf{C}_f(x, y, \theta)$.
- $\Leftrightarrow$ Convexity of $\{(x,t) | t \geq f(x), x \in \mathrm{dom} f\}$ (epigraph)
- When f is differentiable, an equivalent definition is

$$\mathbf{D}_f(y, x) \triangleq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq 0, \forall x, y \in \mathrm{dom} f$$

- We often consider f as an extended convex function on $\mathbb{R}^n$: $\tilde{f}(x) = f(x)$ if $x \in \operatorname{dom} f$ and $+\infty$ otherwise
- Given a convex set A, say $A = \{\|x\| \leq 1\}$, the indicator function, $\iota_A(x) = 0$ if $x \in A$ and $+\infty$ o/w, is convex
  - Constrained objective $\rightarrow$ penalized objective
- Jensen's inequality: Let X be a random variable, f a convex function and $X \in \operatorname{dom} f$, then $f(\mathbb{E}X) \leq \mathbb{E}f(X)$

Strict/strong convexity

- Assume f is convex and differentiable. Then
  - f is strictly convex means $\mathbf{D}_f(y,x) > 0$ for $x \neq y$
  - f is α-strongly convex with $\alpha > 0$ means $\mathbf{D}_f(y,x) \geq \alpha \mathbf{D}_2(y,x) = \alpha\|y-x\|^2/2$
- We do not have to assume differentiability if using $\mathbf{C}_f$. [Strict: $\mathbf{C}_f(x,y,\theta) > 0 \ \forall x \neq y$; strong: $\mathbf{C}_f \geq \alpha \mathbf{C}_2$]
- When $f \in \mathcal{C}^{(2)}$, convexity $\Longleftrightarrow \nabla^2 f(x) \succeq 0$
  - Strong convexity $\Leftrightarrow \nabla^2 f(x) \succeq \alpha I$
  - Strict convexity is *implied* by $\nabla^2 f(x) > 0$

Optimality and convexity

- Let x^o be a (local) minimizer of $f(x)$ with f differentiable. Then $\langle \nabla f(x^o), x - x^o \rangle \geq 0$, $\forall x$ (feasible)
- It follows that $f(x) - f(x^o) \geq \mathbf{D}_f(x, x^o)$ $\forall$ feasible x
  - If f is convex, x^o is a global minimizer; if f is strictly convex, x^o is unique! ($\mathbf{D}_f(y, x) = 0 \rightarrow x = y$)
- Q: Does lasso $\min \|y - X\beta\|_2^2/2 + \lambda\|\beta\|_1$ always lead to a unique solution? How about graphical lasso?

Examples

- $\exp(ax)$, $|x|^a$ with $a \geq 1$, $-\log x$, $x \log x$
- Norm functions
- The max function $f(x) = \max_i x_i$
- $\log \sum \exp(a_i x_i)$ (multinomial)
  - Logistic regression: $-y^T X\beta + \sum \log(1 + \exp(\tilde{x}_i^T \beta))$
- $-\log\det(X)$ $(X \succeq 0)$
  - It suffices to show $-\log\det(Z + tV)$ is convex in t
- Question: Are the last two strictly convex?

‘Convex’ operations

- If f_i are convex, so is $\sum a_i f_i$ as long as $a_i \geq 0$
- If f is convex, $g(x) = f(Ax + b)$ is convex
- If f_i are convex, so is $f(x) = \max_i f_i(x)$. In fact, $f(\cdot, l)$ is convex, $\forall l \in L \Rightarrow \sup_{l \in L} f(x, l)$ is convex
- If h is convex and (its extension) is increasing, and g is convex, then $f(x) = h(g(x))$ is convex
  - Vector: $h(g(x) = h(g_1(x), \ldots, g_n(x))$ is convex if g_i are convex, h is convex and increasing in each argument
- If $f(x, y)$ is (jointly) convex, and C is a nonempty convex set, then $g(x) = \inf_{y \in C} f(x, y)$ is convex

Examples

- If $f(x)$ is convex, $f(-x)$ is convex (& $-f(x)$ is concave)
- $\log(\sum_k \exp(g_k(x)))$ is convex in x, if g_k are convex

- Courant-Fischer **minimax** theorem states that the k-th largest eigenvalue λ_i of a symmetric $A \in \mathbf{S}^n$ is

$$\begin{aligned}\lambda_k(A) &= \sup_{V:\dim V = k} \inf_{v \in V:\|v\|_2=1} v^T A v \\ &= \inf_{\dim V = n-k+1} \sup_{v \in V:\|v\|_2=1} v^T A v\end{aligned}$$

- Intuition: $k=2$: $V = V_{\lambda_1} \oplus V_{\lambda_2}, V_{\lambda_2} \oplus V_{\lambda_3}, V_{\lambda_1} \oplus V_{\lambda_n}, \ldots$
- When $k=1$, $\lambda_1(A) = \sup_{v:\|v\|_2=1} v^T A v$ is a convex function. Indeed $\lambda_1(A)$ gives the matrix norm $\|A\|_2$

- Similarly, we know $\lambda_n(A)$ is concave in A (and any induced matrix-norm is convex)
- A related fact (where $\mathcal{P}_V$ is the projection onto V)

$$\lambda_1(A) + \cdots + \lambda_k(A) = \sup_{\dim(V)=k} tr(A\mathcal{P}_V)$$

$$\lambda_n(A) + \cdots + \lambda_{n-k+1}(A) = \inf_{\dim(V)=k} tr(A\mathcal{P}_V)$$

  $\Rightarrow \lambda_1(A) + \cdots + \lambda_k(A)$ is convex given any $1 \leq k \leq n$
- So is $w_1\lambda_1(A) + \cdots + w_k\lambda_k(A)$ if $w_1 \geq \cdots \geq w_k \geq 0$
- A special case: A is diagonal. Extension: $A \in \mathbb{R}^{n\times p}$?

Sorted ℓ_1 norm for high-dimensional **inference**

- Benjamini-Hochberg (BH) is widely used in multiple testing and controls the FDR level q
- Interestingly, BH can be characterized from an optimization perspective (Abramovich et al 13)
- In the regression setting, let $\lambda_j = \sigma\Phi^{-1}(1 - jq/2n)$ $(1 \le j \le p)$. SLOPE minimizes the following objective

$$\frac{1}{2}\|y - X\beta\|_2^2 + \sum \lambda_j |\beta|_{(j)}$$

- Can you see the convex relaxation? Q: Study how SNR and dependencies affect the power; unknown scale.

Conjugate

- Given $f : \mathbb{R}^n \to \mathbb{R}$, define its conjugate

$$f^*(y) = \sup_{x \in \mathrm{dom} f} \{y^T x - f(x)\}$$

  where $y \in \mathbb{R}^n : y^T x - f(x) < +\infty$ (domain!)
- f^* is convex and closed (whether or not f is convex)
  - Closedness refers to the epigraph
- Surely it has a close connection to Lagrangian, and we have $\langle x, y \rangle \leq f(x) + f^*(y)$ (Fenchel's inequality)

- Examples:
  - Support function of A: given A, $\iota_A^*(y) = \sup_{x \in A} \langle y, x \rangle$
  - $f(x) = x^T Q x / 2$ with Q pd $\Rightarrow f^*(y) = y^T Q^{-1} y / 2$
  - $f(x) = ax + b \Rightarrow f^*(y) = -b + \iota_{\{y=a\}}$
  - $f(x) = \lambda \|x\| (\lambda \geq 0) \Rightarrow f^*(y) = \iota_{\|y\|_* \leq \lambda}$. This is due to Holder's inequality $\langle x, y \rangle \leq \|x\| \|y\|_*$
- Recognizing the conjugate can often facilitate the derivation of the dual problem

Entropy functions

- Let $f(x) = \log \sum_1^n \exp(x_i)$. Then $f^*(y) = \sum y_i \log y_i$, where $y \in \mathbb{R}^n : y_i \geq 0$ and $1^T y = 1$ (negative entropy)
- **KL divergence**: $D(p, q) = \sum p_i \log(p_i / q_i)$ for two discrete probabilities p, q (also called relative entropy)
    - Finiteness needs $q_i = 0 \to p_i = 0$ and $0 \log 0 := 0$
    - Un-normalized: $D(p, q) = \sum \{p_i \log(p_i/q_i) - p_i + q_i\}$, the Bregman divergence of $\sum t_i \log t_i$
- $D(p, q) = -\sum p_i \log q_i + \sum p_i \log p_i = H(p, q) - H(p)$ (cross-entropy minus entropy) and is convex (jointly)

- $\log\{1/P(A)\}$: information or surprisal of event A
  - No surprise, no information
- Entropy tells the average information of a r.v., and offers a measure of <u>uncertainty</u> or <u>disorder</u> (diversity)
- The principle of maximum entropy: Find a distribution satisfying all given constraints (but no more)
  - No **additional** knowledge is to be assumed (no bias)
  - Hence we would like to **maximize** the (remaining) uncertainty or information of the distribution

Maximum entropy

- One rolled a die many times and got an average of 5
- $5 > 3.5$.. The die might not be fair. Then what could be a reasonable estimate of the distribution of X?
- Formulate the optimization problem as

$$\min_p \sum p_i \log p_i \text{ s.t. } p_i \geq 0, \sum p_i = 1, \sum i p_i = 5$$

  or $\min_p D(\{p_i\}, \{1/6\})$ s.t. $p_i \geq 0, \sum p_i = 1, \sum i p_i = 5$
- The prior pmf $\{1/6\}$ can be changed. Constraints extend to $\mathbb{E} f_k(X) = 0$, $1 \leq k \leq K$ (with f_k known)

Most uncertain(informative) vs. most probable

- Constrained multinomial MLE:

$$\max \Pi p_i \text{ s.t. } p_i \geq 0, \sum p_i = 1, \sum i p_i = 5$$

- **Empirical likelihood**: Similarly, we can add a baseline to minimize $-2\log(\Pi p_i/\Pi(1/6)) = 2\sum\{(1/6)\log(1/6) - (1/6)\log p_i\} \propto D(\{1/6\}, \{p_i\})$, subject to the same linear constraints (also convex)
- There is a large body of literature studying these optimization problems and the solutions' asymptotics
- The Cressie-Read (CR) family of divergence measures

Constrained entropy & Poisson GLM

- Let μ_i be the unknowns ($i = 1, \dots, n$), $q_i \geq 0$ be the prior, x_j ($j = 1, \dots, p$) be the features. Consider

$$\min_{\mu} D(\mu, q) = \sum \mu_i \log(\mu_i / q_i) - \mu_i + q_i$$
$$\text{s.t. } x_j^T \mu = \alpha_j, 1 \leq j \leq p, \mu_i \geq 0$$

- From the **dual**, it is equivalent to (non-rigorous for $q_i = 0$)

$$\min_{\beta} \langle q, \exp(X\beta) \rangle - \langle \alpha, \beta \rangle$$

which is the Poisson MLE when $\alpha = X^T y, q = 1$

- Note the difference and relation between μ, q, y
  - A good example: iterative proportional scaling
- We can extend the result to derive distributions in the *exponential* family (by varying the support and moment constraints)
- Another note: "Least informative" priors refer to those containing least information in the constraints (but have maximum entropy or information)

Log-concave functions

- $f(x) > 0$ and $-\log f$ is convex. (It is convenient to use the extended-value function here)
- Assume $\text{vec}(X) \sim \mathcal{N}(0, \Sigma_p \otimes I_n)$. Then $W = X^T X \sim W_p(\Sigma, n)$ (**Wishart**); its density is log-concave in W

$$\propto (\det W)^{(n-p-1)/2} (\det \Sigma)^{-n/2} \exp(-\frac{1}{2} tr(\Sigma^{-1} W))$$

- Also, the reparametrization $\Omega = \Sigma^{-1}$ ensures the log-concavity of the **likelihood** in Ω

- Statisticians frequently encounter log-concave densities/distributions
  - Gibbs random fields, Monte Carlo, majorization, etc.
- A log-concave density is sub-exponential and unimodal
- The product of two log-concave functions is log-concave
- A nice & deep result: If $f(\cdot, \cdot)$ is (jointly) log-concave, then $g(x) = \int f(x, y)\, \mathrm{d}y$ is also log-concave

- Therefore, all marginal densities of a log-concave density are necessarily log-concave
- Also, log-concavity is closed under convolution $(f * g)(x) = \int f(x-y)g(y)\,\mathrm{d}y$
- The distribution of a log-concave density is log concave–$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \int 1_{\geq 0}(x-t)f(t)\,\mathrm{d}t$

Moment/cumulant generating functions

- Due to the composition rules, the sum of log-convex functions is still log-convex (so is their product)
- So $M_X(t) := \mathbb{E}_X \exp\langle t, X\rangle = \int \exp\langle t, X\rangle \,\mathrm{d}F(x)$ is always log-convex!
- Under some regularity conditions, $\nabla M(0) = [\,\mathbb{E}_X(\exp\langle t, X\rangle)']_{t=0} = \mathbb{E}X$, $\nabla^2 M(0) = [\,\mathbb{E}_X \nabla(\exp\langle t, X\rangle X^T)]_{t=0} = \mathbb{E}[XX^T]$
- $m_X(t) = \log M_X(t)$ is convex and $\nabla \log M(0) = \mathbb{E}X$, $\nabla^2 \log M(0) = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T]$
- We will define and study the derivatives later

- Consider bounding $\mathbb{P}[X \in A]$ (A may not be convex)
- Similar to the proof of Markov inequality,

$$\mathbb{P}[X \in A] = \mathbb{E} 1_A(X) \leq \mathbb{E} \exp(\langle t, X \rangle) + \mu),$$

  where $\mu : \mu + \langle t, x \rangle \geq 0,\ \forall x \in A$
- We can solve a convex problem to get a bound, since

$$\begin{aligned} \mathbb{P}[X \in A] &\leq \exp\{\log \mathbb{E} \exp(\langle t, X \rangle) + \sup_{x \in A}(-\langle x, t \rangle)\} \\ &= \exp(m_X(t) + \iota_A^*(-t)) \end{aligned}$$

  where ι_A^* is the support function (conjugate) of A

Generalized convexity

- Given a proper cone $K \subseteq \mathbb{R}^m$ with associated inequality $\preceq_K$, $f : \mathbb{R}^n \to \mathbb{R}^m$ is K-convex if

$$f(\theta x + (1-\theta)y) \preceq_K \theta f(x) + (1-\theta) f(y), \forall x, y, \theta \in [0, 1]$$

- Matrix convexity for $f : \mathbb{R}^n \to \mathbf{S}^m$: $\mathbf{C}_f(x, y, \theta) \succeq 0$, where $\mathbf{S}^m = \{X \in \mathbb{R}^{m \times m} : X = X^T\}$
- *Generalized* inequalities are used on matrix functions

Generalized inequalities

- Let $K \in \mathbb{R}^n$ be a proper cone satisfying (a) cone: for any $x \in C, \theta \geq 0$, $\theta x \in C$; (b) convex; (c) closed; (d) solid ($\mathrm{int} K \neq \phi$); (e) pointed (containing no line)
- $K = \mathbb{R}^n_+$. Norm cones: $C = \{(x, t) : \|x\| \leq t\} \in \mathbb{R}^{n+1}$. Positive semidefinite cone: $\mathbf{S}^n_+ = \{X \in \mathbf{S}^n : X \succeq 0\}$
- K defines a partial ordering on $\mathbb{R}^n$: $y \succeq_K 0 \Leftrightarrow y \in K$, $x \preceq_K y \Leftrightarrow y - x \in K$, $x \prec_K y \Leftrightarrow y - x \in \mathrm{int} K$.
- In particular, for two symmetric matrices $X, Y \in \mathbf{S}^n$, $X \succeq Y$ (associated with $\mathbf{S}^n_+$) means $X - Y$ is psd

Convex optimization

- Consider the optimization problem:

  $$\min f_0(x) \text{ s.t. } f_i(x) \leq 0, 1 \leq i \leq m,\ h_j(x) = 0, 1 \leq j \leq p$$

- Feasibility problems: $f_0 = 0$. So the optimal value $p^\star \triangleq \inf\{f_0(x) : f_i(x) \leq 0, 1 \leq i \leq m,\ h_j(x) = 0, 1 \leq j \leq p\}$ is either 0 or $+\infty$ (if the feasible set is empty)
- Convex optimization: $f_0, f_1, \ldots, f_m$ are convex and the equality constraints are **affine**: $a_j^T x = b_j$, $1 \leq j \leq p$

Examples of convex programming

- Linear program: $\min c^T x + d$ s.t. $Gx \preceq h, Ax = b$.
  Standard form LP: $\min c^T x$ s.t. $Ax = b, x \succeq 0$
- Quadratic program (QP): $\min x^T Px/2 + q^T x + r$ s.t. $Gx \preceq h, Ax = b$, where P is psd ($P \in S^n_+$)
- Quadratically constrained quadratic program (QCQP): $\min x^T Px/2 + q^T x + r$ s.t. $x^T P_i x/2 + q_i^T x + r_i \leq 0, 1 \leq i \leq m, Ax = b$, where P, P_i are psd
- Second-order cone program (SOCP): $\min f^T x$ s.t. $\|A_i x + b_i\|_2 \leq c_i^T x + d_i, 1 \leq i \leq m, Fx = g$.
  - Note the second-order (norm) cone in form of $\|x\|_2 \leq t$

Semidefinite programming

- Convex optimization with generalized inequality constraints: $\min f_0(x)$ s.t. $f_i(x) \leq_{K_i} 0$, $1 \leq i \leq m$, $Ax = b$, where K_i are proper cones, f_0, f_i are convex
  - Ordinary convex programming: $K_i = \mathbf{R}_+$
- Conic-form problems: $\min c^T x$ s.t.$Fx + g \preceq_K 0, Ax = b$
- SDP: $K = \mathbf{S}^k_+$, i.e., for $x \in \mathbb{R}^n$, $F_i, G \in \mathbf{S}^k$, $A \in \mathbb{R}^{\times n}$

$$\min c^T x \text{ s.t. } x_1 F_1 + \cdots + x_n F_n + G \preceq 0, Ax = b$$

- Note the **linear** matrix inequality (& linear objective)

- Standard form SDP (with a matrix variable):

$$\min tr(CX) \text{ s.t. } tr(A_iX) = b_i (1 \le i \le p), X \succeq 0$$

  where X, C, A_i are symmetric matrices of the same size
- **LP $\subset$ QP $\subset$ QCQP $\subset$ SOCP $\subset$ SDP**
  - QCQP $\subset$ SOCP: Convex quadratic constraint can be written as SOC; $\min t$ s.t. $x^T Px/2 + 0 \cdot t^2 + q^T x - t + r \le 0, x^T P_i x/2 + q_i^T x + r_i \le 0, 1 \le i \le m, Ax = b$
  - SOCP $\subset$ SDP: <u>2nd-order</u> constraints can be converted to LMI using the (generalized) Schur complement:

$$\|x\|_2 \le t \Leftrightarrow t - x^T t^+ x \succeq 0, t \succeq 0, (1-tt^+)x = 0 \Leftrightarrow \begin{bmatrix} t & x^T \\ x & tI \end{bmatrix} \succeq 0$$

- SDP can be efficiently solved by say interior point methods, and has very good software support
  - SeDuMi, SDPT3, SDPA, ...
  - Also, check CVX, CVXOPT
- However, the guaranteed polynomial-complexity is still prohibitive on large (or even moderate) problems

Robust programming

- Let's start with an LP: $\min c^T x$ s.t. $a_i^T x \leq b_i, 1 \leq i \leq n$
- Assume a_i are Gaussian **random** vectors $\mathcal{N}(\bar{a}_i, \Sigma_i)$ and each constraint holds with probability at least η

$$\min c^T x \text{ s.t. } \mathbb{P}(a_i^T x \leq b_i) \geq \eta, 1 \leq i \leq n$$

- Since $\mathbb{P}(z \leq b) \geq \eta \Leftrightarrow b \geq F_z^{-1}(\eta)$, we get an SOCP

$$\min c^T x \text{ s.t. } \bar{a}_i^T x + \Phi^{-1}(\eta) \|\Sigma_i^{1/2} x\|_2 \leq b_i, 1 \leq i \leq n$$

A latent variable model for classification

- Let $y_i = \begin{cases} 1, & \tilde{x}_i^T\beta + \epsilon_i > 0 \\ 0, & \tilde{x}_i^T\beta + \epsilon_i \leq 0 \end{cases}$, where $\epsilon_i \overset{iid}{\sim} F$
- Assume F has a **log-concave** density f
- y_i is Bernoulli: $\pi_i = 1 - F(-\tilde{x}_i^T\beta) = F((-\tilde{x}_i^T\beta, +\infty))$
- $L(\beta) = -\sum \log(1-\pi_i) - \sum y_i \log(\pi_i/(1-\pi_i)) =$
  $-\sum y_i \log F(-\tilde{x}_i^T\beta) - \sum(1-y_i)\log(F((-\tilde{x}_i^T\beta, +\infty))$
- L is always convex in β (why?)
- A perhaps more convenient way: $y_i = 1_{\tilde{x}_i^T\beta \geq \varepsilon_i}$, $\varepsilon_i \overset{iid}{\sim} F$

Some special cases (with different tails & symmetry)

- $F = \Phi$ gives Probit models ($F^{-1}(\pi_i) = \tilde{x}_i^T\beta$)
- Assuming a logistic distribution for ϵ_i (symmetric):

$$f(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}, \;\; F(x) = \frac{1}{1+\exp(-x)}$$

  we get the logistic regression
- Gumbel distribution: $F(t) = \exp(-\exp(-t))$, $f(t) = \exp(-t - \exp(-t))$ (nonsymmetric but log-concave)$\rightarrow$ negative log-log link $g = -\log(-\log(\pi))$
- Choice? Relate the link function to ε_i's distribution!

Max cut

- let G be an undirected graph with n nodes and n^2 weights $a_{ij} \geq 0$ placed on the edges $(a_{ij} = a_{ji})$
- The problem is to find a partition $S \cup S^C = [n]$ to maximize the sum of the weights of the 'crossing' edges
- Let $S = \{i \in [n] : x_i = 1\}$ and $S^c = \{i \in [n] : x_i = -1\}$.

$$\max_{x \in \{-1,1\}^n} \frac{1}{2} \sum_{i,j} a_{ij} 1_{x_i x_j = -1} = \frac{1}{4} \sum a_{ij}(1 - x_i x_j)$$

- Equivalent to $\min_{x \in \{-1,1\}^n} x^T A x$.

- The IP is well known to be NP-hard. Convex relation?
  - Replace the constraints by $-1 \leq x_i \leq 1$ (and use $A + \lambda I$ in place of A? $\|x\|_2^2 = n$ on $\{-1, 1\}^n$)
- To make a better one, introduce $X = xx^T$ which satisfies $X \succeq 0$, $X_{i,i} = 1$, and $\text{rank}(X) = 1$. Consider

$$\min_{X \in \mathbf{S}^n} \langle A, X \rangle \text{ s.t. } X_{i,i} = 1, X \succeq 0$$

- The SDP gives impressive bounds on the optimal value. Need an extra <u>rounding</u> to get an approximate solution

Nuclear norm minimization

- Let's consider a matrix completion problem

$$\min_{X \in \mathbb{R}^{n \times m}} \operatorname{rank}(X) \text{ s.t. } \mathcal{A}(X) = b$$

  where $\mathcal{A}(\cdot)$ denotes a linear mapping
- Enforcing low rank is natural and effective
  - Robust PCA, video inpainting, recommender systems
- A convex relaxation can be made with nuclear norm

$$\min_{X \in \mathbb{R}^{n \times m}} \|X\|_* \text{ s.t. } \mathcal{A}(X) = b$$

- How to deal with the nondifferentiable objective?

- Fazel (02): $\|X\|_* \leq t \iff tr(W_1)/2 + tr(W_2)/2 \leq t$ for some W_i satisfying $\begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix} \succeq 0$. (Use SDP duals!)
- Hence we get an **SDP** as follows

$$\min_{X \in \mathbb{R}^{n \times m}, W_1 \in \mathbf{S}^n, W_2 \in \mathbf{S}^m} tr(W_1)/2 + tr(W_2)/2$$

$$\text{s.t. } \mathcal{A}(X) = b, \begin{bmatrix} W_1 & X \\ X^T & W_2 \end{bmatrix} \succeq 0$$

- Other ways exist to solve it (e.g., **proximal** methods)

A sparse PCA

- Recall that given $X = UDV^T$, $XV_rV_r^T$ gives the best rank-r approximation to X in the sense of F-norm
- We can minimize $\|X - X\mathbf{P}\|_F^2$ over all rank-r projection matrices, or equivalently, $\max_{\mathbf{P}\in\mathcal{P}^r}\langle\Sigma, \mathbf{P}\rangle$
- Here, $\Sigma = X^TX$, $\mathcal{P}^r = \{P : P^2 = P, P^T = P, r(P) = r\}$
- A formulation of the sparse PCA problem is

$$\max_{\mathbf{P}\in\mathcal{P}^r}\langle\Sigma, \mathbf{P}\rangle - \lambda\|\operatorname{vec}(\mathbf{P})\|_1$$

- A convex relation (**Fantope** of order r, Vu et al 13):

$$\mathcal{F}^r = \{P : 0 \preceq P \preceq I, tr(P) = r\}$$

- The resulting problem is an SDP. (Of course, we can solve it using other methods, such as ADMM.)
- Do you think if pursuing sparsity in $\mathbf{P}$ makes sense? Later we will introduce other forms of sparse PCA.