Matrix Algebra and Optimization for Statistics and Machine Learning

Yiyuan She

Department of Statistics, Florida State University

- Duality and constrained optimization

The primal problem

- Consider a constrained optimization problem:

$$\min_{x \in \mathbb{D}} f_0(x)$$
$$\text{s.t. } f_i(x) \leq 0, 1 \leq i \leq m, h_i(x) = 0, 1 \leq i \leq p$$

- $\mathbb{D} \subset \mathbb{R}^n$: $\mathrm{dom}(f_0) \cap \mathrm{dom}(f_1) \cap \cdots \cap \mathrm{dom}(h_p)$
- **Convex** programming: f_0, f_i are all convex, h_i affine

Conversions

- f_0: We can change the objective to a constraint $f_0(x) \le t$ and minimize t
- f_i: we can change an inequality constraint to an equality one: $f_i(x) + s_i = 0,\ s_i \ge 0$ (s_i: slack variables)
  - $s_i \ge 0$: barrier/proximity. Alternatively, $f_i(x) = -s_i^2$
- h_i: $\pm h_i \le 0$. Sometimes we introduce additional equality constraints to decouple (e.g., $\min f(x) + g(x)$)

Lagrangian

- Lagrangian: $L(x, \lambda, \nu) = f_0(x) + \langle \lambda, \vec{f}(x) \rangle + \langle \nu, \vec{h}(x) \rangle$, where $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}^m$, $\nu \in \mathbb{R}^p$, $\lambda_i \geq 0$
  - Due to the implicit constraints, the domain for x may be smaller than $\mathbb{R}^n$: $\mathcal{D} = (\cap_0^m \operatorname{dom} f_i) \cap (\cap_1^p \operatorname{dom} h_i)$
  - *Primal feasibility*: $x \in \mathcal{D}$, $f_i(x) \leq 0$, $h_i(x) = 0$
- $\lambda_i (1 \leq i \leq m)$, $\nu_i (1 \leq i \leq p)$: Lagrangian multipliers or dual variables
- Lagrange dual function: $g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$
- Note the infimum could result in $-\infty$

- *Dual feasibility* of (λ, ν): $\lambda \succeq 0 (\geq 0)$, $g(\lambda, \nu) > -\infty$
- Perhaps interestingly, the Lagrange dual problem

$$\max_{\lambda,\nu} g(\lambda, \nu) \text{ s.t. } \lambda \geq 0 \text{ (and } g(\lambda, \nu) > -\infty)$$

often yields a solution to achieve the optimum of the original optimization problem in the **convex** case

Why turn to the dual?

- The dual problem (always convex) may provide useful information and offer a lot of ease in optimization
- Dimensions for the primal and dual: n vs. $m + p$
- In evaluating the dual function, one freely considers $x \in \mathcal{D}$ without any further restrictions!

Duality gap

- Let $x^\star$ be a globally optimization solution to the primal problem, and $p^\star = f_0(x^\star)$
- Then, it is clear that $g(\lambda, \nu) \leq f_0(x^\star) = p^\star$.
- Let $d^\star$ be the optimal function value of the dual problem. Then $d^\star \leq p^\star$, referred to as the weak duality.
- The (optimal) duality gap is defined as $p^\star - d^\star$, which is crucial for analysis and implementation

Strong duality

- In **convex** programming, the **strong** duality

$$d^\star = p^\star,$$

  is implied by strict feasibility (Slater's condition)
- Concretely, for min $f_0(x)$ s.t. $f_i \le 0, Ax = b$, strict feasibility means $\exists x \in \mathrm{relint}\mathcal{D}$ s.t. $f_i(x) < 0, Ax = b$
- Weak Slater: "<" for non-affine inequalities only $\Rightarrow$SD
  - Applies to ordinary convex programming (not LMI)
- Other constraint qualifications (CQ) exist

Example: LP

- LP in standard form: $\min c^T x$ s.t. $Ax = b, x \geq 0$
- $L(x, \lambda, \nu) = c^T x - \lambda^T x + \nu^T (Ax - b)$
- The Lagrange dual is given by

$$
\begin{aligned}
g(\lambda, \nu) &= \inf_x (c + A^T \nu - \lambda)^T x - b^T \nu \\
&= \begin{cases} -b^T \nu, & A^T \nu - \lambda + c = 0 \\ -\infty, & \text{o/w} \end{cases}
\end{aligned}
$$

- The dual problem: $\max_{\lambda \geq 0, \nu} -b^T \nu$ s.t. $A^T \nu + c = \lambda$ or $\max_\nu -b^T \nu$ s.t. $A^T \nu + c \geq 0$

Example: entropy maximization

- Consider $\min \sum x_i \log x_i$ s.t. $Ax \preceq b, 1^T x = 1$ which is a special case of $\min f_0(x)$ s.t. $Ax \preceq b, Cx = d$
- With linear constraints, we can use conjugate to obtain

$$\begin{aligned} g(\lambda, \nu) &= \inf f_0(x) + \lambda^T (Ax - b) + \nu^T (Cx - d) \\ &= -b^T \lambda - d^T \nu + \inf_x \{ f_0(x) + (A^T \lambda + C^T \nu)^T x \} \\ &= -b^T \lambda - d^T \nu - \sup_x \{ (-A^T \lambda - C^T \nu)^T x - f_0(x) \} \\ &= -b^T \lambda - d^T \nu - f_0^*(-A^T \lambda - C^T \nu) \end{aligned}$$

- $\mathrm{dom} g = \{ (\lambda, \mu) : -A^T \lambda - C^T \nu \in \mathrm{dom} f_0^* \}$ (no $-\infty$!)

- For the negentropy function, $f_0^*(y) = \sum \exp(y_i - 1)$
- The dual problem is a Poisson-type problem

$$\max_{\lambda,\nu} -b^T\lambda - \nu - \exp(-\nu - 1)\langle 1, \exp(A^T\lambda)\rangle \text{ s.t. } \lambda \succeq 0$$

- Evaluating ν gives a multinomial-type problem

$$\max_{\lambda \succeq 0} -b^T\lambda - \log\langle 1, \exp(A^T\lambda)\rangle$$

- Note the dimension change. **EL** uses the same trick.
- Weak Slater's condition: $\exists x \succ 0$ with $Ax \preceq b, 1^T x = 0$.

Example: generalized lasso

- Joint regularization imposes multiple penalties on β
  - Fused lasso: $P(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \sum |\beta_{j+1} - \beta_j|$
  - Spare group lasso: $P(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \sum \|\beta^k\|_2$
  - Clustered lasso: $P(\beta) = \lambda_1 \|\beta\|_1 + \lambda_2 \sum_{j \neq j'} |\beta_j - \beta_{j'}|$
- Let's begin with $\min l(\beta) + P(T\beta)$, which can be rephrased as $\min_{\beta,\gamma} l(\beta) + P(\gamma)$ s.t. $T\beta = \gamma$
- $g(\nu) = \inf_{\beta,\gamma} l(\beta) + P(\gamma) + \nu^T(T\beta - \gamma) = -\sup_{\gamma}\{\nu^T\gamma - P(\gamma)\} - \sup_{\beta}\{\nu^T(-T\beta) - l(\beta)\} = -l^*(-T^T\nu) - P^*(\nu)$

- Example: $l(\beta) = \|\beta - y\|_2^2/2$ and $P(\gamma) = \lambda\|\gamma\|_1$.
- From $l^*(\eta) = \|\eta + y\|_2^2/2 - \|y\|_2^2/2$, $P^*(\eta) = \iota_{\|\eta\|_\infty \leq \lambda}$, the dual problem is given by

$$\max_{\nu:\|\nu\|_\infty \leq \lambda} -\frac{1}{2}\|T^T\nu - y\|_2^2 + \frac{1}{2}\|y\|_2^2$$
$$= \frac{1}{2}\|y\|_2^2 - \min_{\nu:\|\nu\|_\infty \leq \lambda} \frac{1}{2}\|T^T\nu - y\|_2^2$$

- Example: $l(\Omega) = \langle \Sigma, \Omega \rangle - \log\det\Omega$, $P(\Omega) = \lambda\|\Omega\|_1$, $\Omega \in \mathbf{S}_{++}^n$. Recall $(-\log\det)^*(\tilde{\Omega}) = -\log\det(-\tilde{\Omega}) - n$
- With $T = I$, the dual problem is $n - \min_{\tilde{\Omega}-\Sigma \succ 0, \|\tilde{\Omega}\|_{\max} \leq \lambda} -\log\det(\tilde{\Omega} - \Sigma)$ or $n - \min_{\|\tilde{\Omega}-\Sigma\|_{\max} \leq \lambda, \tilde{\Omega} \succ 0} -\log\det(\tilde{\Omega})$

- Next, let $l(\eta) = \|\eta - y\|_2^2/2$, $\eta = X\beta$, and $P(\gamma) = \lambda\|\gamma\|_1$, $\gamma = T\beta$. Similarly, the dual problem of $\min_{\beta,\gamma,\eta} l(\eta) + P(\gamma)$ s.t. $T\beta = \gamma, X\beta = \eta$ is

$$\max_{\lambda,\nu} -l^*(\nu) - P^*(\mu) \text{ s.t. } T^T\mu + X^T\nu = 0$$

  or $\frac{1}{2}\|y\|_2^2 - \min \frac{1}{2}\|\nu + y\|_2^2$ s.t. $T^T\mu = -X^T\nu, \|\mu\|_\infty \leq \lambda$
- Lasso ($T = I$): $\min_{\nu\in\mathbb{R}^n} \|\nu + y\|_2^2/2$ s.t. $\|X^T\nu\|_\infty \leq \lambda$ or $\min_{\nu\in\mathbb{R}^n} \|\nu\|_2^2/2$ s.t. $\|X^T(y-\nu)\|_\infty \leq \lambda$. (Compare it with the Dantzig selector.)

- Last, consider $\min \|y-\beta\|_2^2/2 + P_1(\beta) + P_2(\beta)$ (after linearization). We could rewrite it as

$$\min \|y-\beta\|_2^2/2 + P_1(\beta_1) + P_2(\beta_2), \text{ s.t. } \beta = \beta_1, \beta = \beta_2$$

- $L = \|y-\beta\|_2^2/2 + P_1(\beta_1) + P_2(\beta_2) + \mu^T(\beta-\beta_1) + \nu^T(\beta-\beta_2)$
- $g(\mu,\nu) = \|y\|_2^2/2 - \|y-\mu-\nu\|_2^2/2 - P_1^*(\mu) - P_2^*(\nu)$ since $\beta^o(\mu,\nu) = y-\mu-\nu$.

- The dual problem is thus equivalent to

$$\min_{\mu,\nu} \|y - \mu - \nu\|_2^2/2 + P_1^*(\mu) + P_2^*(\nu)$$

  which is way simpler than the primal!
- We could use proximal gradient descent, or BCD which leads to **Dykstra's projections**

Example: nuclear norm optimization

- Low rank matrix estimation is often achieved by $\|B\|_*$
- It is well known $\|B\|_* = \max_X \langle B, X \rangle$ s.t. $\|X\|_2 \leq 1$
- $\|X\|_2 \leq t \Leftrightarrow t^2 I - XX^T \succeq 0, t \geq 0 \Leftrightarrow \begin{bmatrix} tI & X \\ X^T & tI \end{bmatrix} \succeq 0$ (Schur complement)
- Therefore, the dual-norm $\|B\|_*$ optimization is an SDP

$$\max_X \langle B, X \rangle \text{ s.t. } \begin{bmatrix} I & X \\ X^T & I \end{bmatrix} \succeq 0$$

- The Lagrangian with $W \succeq 0$ is

$$L(X,W) = \langle B, X \rangle + \langle \begin{bmatrix} W_1 & W_{12} \\ W_{12}^T & W_2 \end{bmatrix}, \begin{bmatrix} I & X \\ X^T & I \end{bmatrix} \rangle$$
$$= \langle B, X \rangle + tr\{W_1\} + tr\{W_2\} + 2\langle W_{12}, X \rangle$$

- With $W_1 \leftarrow 2W_1, W_2 \leftarrow 2W_2$, the dual SDP is

$$\min_{W_1,W_2} \frac{1}{2}(tr\{W_1\} + tr\{W_2\}) \text{ s.t. } \begin{bmatrix} W_1 & B \\ B^T & W_2 \end{bmatrix} \succeq 0$$

[The sign of B does not matter.]

Saddle-point characterization

- We can write $p^\star = \inf_x \sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu)$ due to

$$\sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu) = \begin{cases} f_0(x), & f_i(x) \leq 0, h_i(x) = 0 \\ +\infty, & \text{o/w} \end{cases}$$

- Strong duality means

$$d^\star = \sup_{\lambda \succeq 0, \nu} \inf_x L(x, \lambda, \nu) = \inf_x \sup_{\lambda \succeq 0, \nu} L(x, \lambda, \nu) = p^\star$$

- [Weak duality naturally holds (for any L)]

- Assume only inequality constraints exist (conversion)
- Finding the primal-dual pair $(x^\star, \lambda^\star)$ amounts to finding a saddle point of L in the sense that

$$L(x^\star, \lambda) \leq L(x^\star, \lambda^\star) \leq L(x, \lambda^\star), \forall \lambda \succeq 0, \forall x$$

- More rigorously, (i) a saddle point $\rightarrow$ primal-dual pair; (ii) convex f_i + Slater $\rightarrow \forall x^\star$, $\exists \lambda \succeq 0$ to make $(x^\star, \lambda)$ a saddle point; (iii) convex & differentiable f_i + weak Slater $\rightarrow \forall x^\star$, $\exists \lambda \succeq 0$ to make $(x^\star, \lambda)$ a saddle point.

Optimality conditions

- We'll define a set of KKT optimality conditions to connect $x^\star$ and $\lambda^\star, \nu^\star$
- **Necessity**: If f_0, f_i, g_i are all differentiable and the strong duality holds, any globally optimal solutions (primal and dual) must satisfy the KKT conditions
- **Sufficiency**: If the problem is convex & differentiable, KKT conditions are also sufficient

KKT Conditions

$$\begin{cases} \text{Stationarity: } \nabla f_0(x^\star) + \sum \lambda_i^\star \nabla f_i(x^\star) + \sum \nu_i^\star \nabla h_i(x^\star) = 0, \\ \text{Dual feasibility: } \lambda_i^\star \geq 0, \\ \text{Complementary slackness: } \lambda_i^\star f_i(x^\star) = 0, \\ \text{Primal feasibility: } f_i(x^\star) \leq 0, h_i(x^\star) = 0. \end{cases}$$

Example: SVM

- Recall SVM for classification ($y_i = \pm 1$)

$$\min_{\beta,\beta_0} \sum_{i=1}^{n} [1 - y_i(x_i^T \beta + \beta_0)]_+ + \frac{\lambda}{2} \|\beta\|_2^2$$

- The hinge loss provides a convex relaxation for the misclassification error loss $1_{y_i f_i < 0}$ with $f_i = x_i^T \beta + \beta_0$
- [Question: Pros and cons of the hinge loss?]
- With $[1 - y_i f_i]_+ \leq \xi_i$, $C = \frac{1}{\lambda}$, redefine the problem as

$$\min_{\beta,\beta_0,\xi} \frac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^{n} \xi_i \text{ s.t. } \xi_i \geq 0, y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \forall i$$

- Introduce Lagrangian multipliers $\mu_i, \alpha_i \geq 0$; we get $L = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i - \sum \alpha_i[y_i(x_i^T\beta + \beta_0) - (1 - \xi_i)] - \sum \mu_i \xi_i$
- The dual problem is given by $\max_{\alpha,\mu} \langle 1, \alpha\rangle - \frac{1}{2}(y \circ \alpha)^T (XX^T)(y \circ \alpha)$ s.t. $\alpha_i \geq 0, \mu_i \geq 0, \alpha_i + \mu_i = C, \sum \alpha_i y_i = 0$ (dual feasible!) which reduces to

$$\max_{0 \leq \alpha_i \leq C, \langle \alpha, y\rangle = 0} g(\alpha) := \langle 1, \alpha\rangle - \frac{1}{2}(y \circ \alpha)^T (XX^T)(y \circ \alpha).$$

The KKT conditions are

$$\begin{cases} \beta & = \sum \alpha_i y_i x_i \\ 0 & = \sum \alpha_i y_i \\ 0 & = C - \mu_i - \alpha_i, \forall i \\ 0 & \leq \alpha_i, \forall i \\ 0 & = \alpha_i [y_i (x_i^T \beta + \beta_0) - (1 - \xi_i)], \forall i \\ 0 & = \mu_i \xi_i, \forall i \\ 0 & \leq \mu_i, \forall i \\ 0 & \leq y_i (x_i^T \beta + \beta_0) - (1 - \xi_i), \forall i, \\ 0 & \leq \xi_i, \forall i \end{cases}$$

- The observations with $\hat{\alpha}_i > 0$ are called support vectors, because $\hat{\beta} = \sum_{i:\hat{\alpha}_i \neq 0} \hat{\alpha}_i y_i x_i$.
- $\hat{\xi}_i > 0 \Rightarrow \hat{\mu}_i = 0 \Rightarrow \hat{\alpha}_i = C$ (while for other support vectors with $\hat{\xi}_i = 0$, we have $0 < \hat{\alpha}_i \leq C$)
- ξ_i: allowances. All samples with $\hat{\xi}_i > 1$ are misclassified in the training data and are support vectors. (Does this make sense?)

Optimization algorithms

- The dual form of the constrained optimization problem can be used to design an algorithm
  - Dual BCD, dual ascent, primal-dual methods, etc.
- In the following, we introduce some algorithms for equality/inequality-constrained optimization

Affine-equality constrained optimization

- Consider the problem of $\min f(x)$ s.t. $Ax = b$
- For simplicity, A has full row rank, $f \in \mathcal{C}^{(2)}$ & convex
- Elimination: $x = A^{+}b + U_{\perp}^{T} z$ where $A^{T} = UDV^{T}$
- $\min_{z} f(A^{+}b + U_{\perp}^{T} z)$ is constraint free, and we can apply Newton (no need of explicit elimination)
- But it may not be efficient $(A^{+}b, U_{\perp})$. **Dual**?

- Dual problem: $\max_\nu g(\nu) = -b^T\nu - f^*(-A^T\nu)$
  - Beware of the implicit constraints
- Since the problem is convex, we can use

$$\nu^{t+1} = \nu^t + \alpha_t(Ax^{t+1} - b),$$

  where $x^{t+1} \in \arg\min_x L(x, \nu^t)$
- If x^{t+1} is unique, it becomes gradient ascent (o/w we only have $Ax^{t+1} - b \in \partial g(\nu^t)$ & it can be very slow)
- If $g \in \mathcal{C}^{(2)}$ (not guaranteed), we can use Newton

Example: empirical likelihood for regression

- Given (y, X), to test the hypothesis $\beta = \beta^0$, EL solves

$$\min_{w \in \mathbb{R}^n} -\sum \log w_i \text{ s.t. } 1^T w = 1, X^T \text{diag}\{w\}(X\beta^0 - y) = 0$$

- Newton's method with equality constraint requires a primal feasible w^0 satisfying the constraints
- Let $r = X\beta^0 - y$. The 2nd constraint is $X^T(r \circ w) = 0$
- $L(w, \nu_0, \nu) = -\sum \log w_i + \nu_0(1^T w - 1) + \langle r \circ (X\nu), w \rangle$

- From $[1/w_i] + \nu_0 1 + r \circ (X\nu) = 0$ and other KKT conditions,

$$-1 + \nu_0 w_i + w_i r_i x_i^T \nu = 0 \Rightarrow \begin{cases} w_i = (\nu_0 + r_i x_i^T \nu)^{-1} \\ \nu_0 = n \end{cases}$$

- The dual problem is equivalent to

$$\max_{\nu \in \mathbb{R}^p} \sum \log(n + r_i x_i^T \nu)$$

- *Implicit* constraints: $r_i x_i^T \nu > -n, \forall i$ (cf. **pseudo-log**)
- Newton's method on the dual also needs a feasible start

Infeasible-start Newton

- This is a primal-dual method.
- The KKT equations (primal/dual feasibility equations)

$$Ax^\star = b, \quad \nabla f(x^\star) + A^T \nu^\star = 0$$

- Using the approximation to $\nabla f(x + \Delta x)$, we can get Newton's direction Δx by solving

$$Ax + A\Delta x = b, \nabla f(x) + \nabla^2 f(x)\Delta x + A^T w = 0$$

- The quadratic system can be solved by

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ w \end{bmatrix} = - \begin{bmatrix} \nabla f(x) \\ Ax - b \end{bmatrix}$$

- [Question: When will the KKT matrix be singular?]
- If x is primal feasible, $Ax - b = 0$ (but it needs not be)
- We only need $x^0 \in \mathcal{D}$!

Inequality constraints & interior-point methods

- Consider $\min f_0(x)$ s.t. $f_i(x) \leq 0$, $h_i(x) = 0$, or $\min f_0 + \sum \iota_{f_i \leq 0}$, $h_i = 0$
- A (primal) logarithmic barrier method: solve

$$\min f_0(x) + \frac{1}{\rho} \sum -\log(-f_i(x)) \text{ s.t. } h_i(x) = 0$$

  for a positive sequence of $\rho = \rho_k \to +\infty$
- The barrier term is monotone, smooth, and convex in f_i, and prevents $f_i(x)$ from getting too close to 0

- [Similarly, we can use $\rho\|\vec{h}(x)\|_2^2$ with $\rho \to +\infty$ to deal with the equality constraints. However, this quadratic penalty method is often not that efficient.]
- In the convex setting (f_i convex and h_i affine), we can call Newton's method for each ρ and use path-following
- The optimization for large ρ may be much difficult

Primal-dual interior-point methods

- The barrier method needs $x^0 \in \mathcal{D}$ satisfying $f_i(x) < 0$
- Primal-dual updates are attractive and efficient
  - This class of methods are extremely popular in convex programming and are standard for solving SDP
- A useful form to derive interior-point methods is to use slack variables: $\min f_0(x)$ s.t. $f_i(x) + s_i = 0$, $h_i(x) = 0$, $s \succeq 0$. The log-barrier problem is then given by

$$\min f_0(x) - \frac{1}{\rho} \sum \log s_i \text{ s.t. } f_i(x) + s_i = 0, h_i(x) = 0$$