Convex Optimization

Prof. Nati Srebro

Lecture 8:
From the Fenchel Conjugate to Matrix Inequalities

Reading: Boyd and Vandenberghe Sections 3.3, 4.6, 5.1.6, 5.2.4, 5.7
Recommended reading on specific applications: 6.2, 8.6

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f_0(x) \\ s.t. \quad & f_i(x) \leq 0 \quad i = 1..m \\ & h_i(x) = 0 \quad j = 1..p \end{aligned}$$

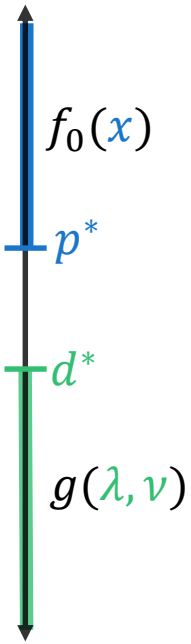
$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^p} \quad & g(\lambda, \nu) \\ s.t. \quad & \lambda_i \geq 0 \end{aligned}$$

$$L\big(x, (\lambda, \nu)\big) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$$

$$g(\lambda, \nu) = \inf_x L\big(x, (\lambda, \nu)\big)$$

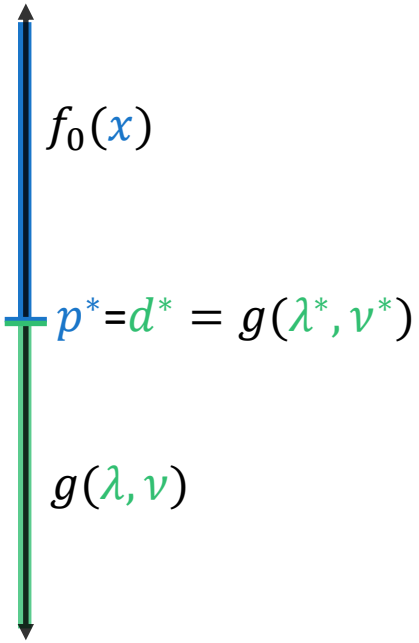
Weak Duality (always holds):

$$d^* \leq p^*$$



Convex + Slater ➔ Strong Duality

$$d^* = p^*$$



Dual of a Function

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & x = 0 \end{aligned}$$

$$\max_{\nu \in \mathbb{R}^n} \quad -f^*(-\nu)$$

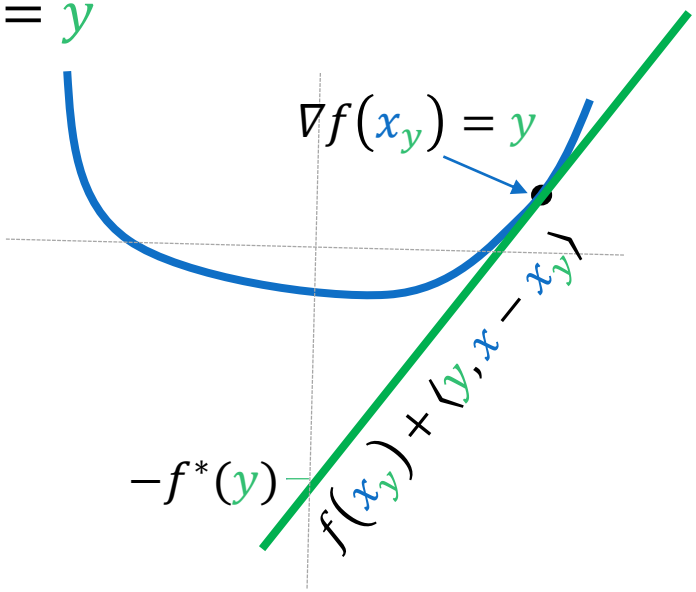
$$g(\nu) = \inf_x L(x, \nu) = \inf_x f(x) + \langle \nu, x \rangle$$

- **Def**: The Fenchel Conjugate of $f: \mathbb{R}^n \to \mathbb{R}$ is a function $f^*: (\mathbb{R}^n)^* \to \mathbb{R}$

$$f^*(y) = \sup_x \langle y, x \rangle - f(x)$$

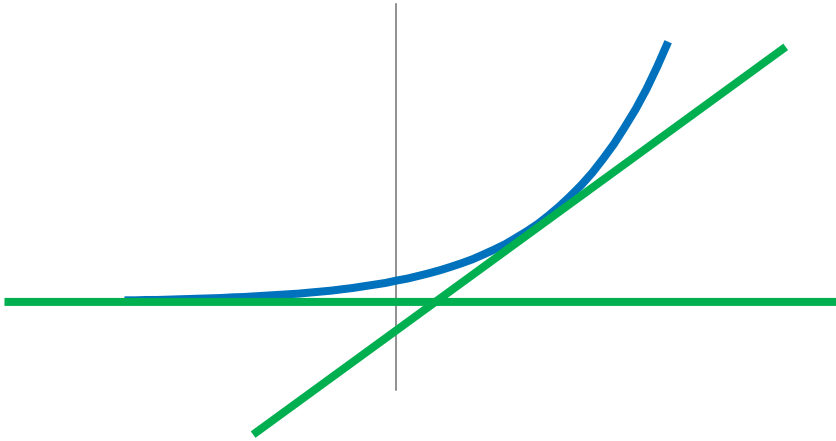
- Claim: $f^*(y) = \langle y, x_y \rangle - f(x_y)$ where $\nabla f(x_y) = y$

Proof: $0 = \nabla_x \big(\langle y, x \rangle - f(x) \big) = y - \nabla f(x)$

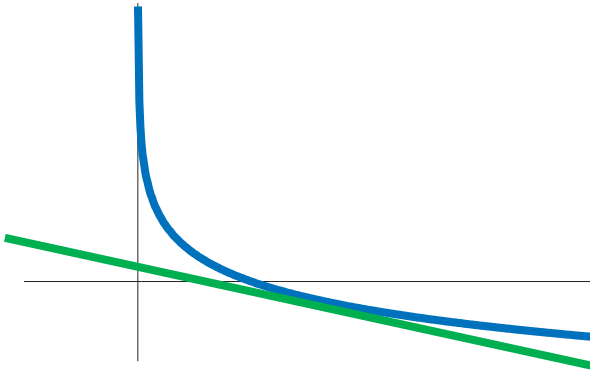


$$f^*(y) \stackrel{\text{def}}{=} \sup_x \langle y, x \rangle - f(x) = \langle y, x_y \rangle - f(x_y) \text{ where } \nabla f(x_y) = y$$

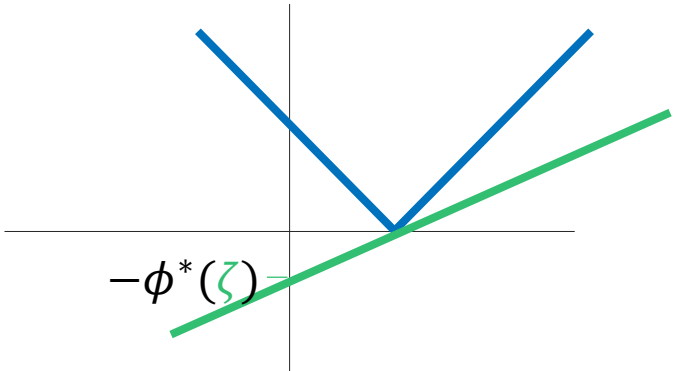
- $f(x) = e^x$
  - $y = f'(x) = e^x \rightarrow x = \log y$
  - $f^*(y) = \begin{cases} y \log y - y & y > 0 \\ 0 & y = 0 \\ \infty & y < 0 \end{cases}$



- $f(x) = -\log x$
  - $y = f'(x) = -\frac{1}{x} \rightarrow x^* = -\frac{1}{y}$ for $y < 0$
  - $f^*(y) = \begin{cases} y\left(\frac{-1}{y}\right) + \log\frac{-1}{y} = \mathbf{-1 - \log(-y)} & y < 0 \\ \infty & y \geq 0 \end{cases}$



- $f(x) = |x - 1|$
  - $y = f'(x) \rightarrow x = 1$ for $-1 \leq y \leq 1$
  - $f^*(y) = \begin{cases} y \cdot 1 - |1 - 1| = \boldsymbol{y} & -1 \leq y \leq 1 \\ \infty & \text{otherwise} \end{cases}$



Optimization with Linear Constraints

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & Ax = b \\ & Gx \leq h \end{aligned}$$

$$\begin{aligned} \max \quad & -\langle \lambda, h \rangle - \langle \nu, b \rangle - f^*(-G^\top \lambda - A^\top \nu) \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned}$$

$$\begin{aligned} g(\lambda, \nu) &= \inf_x f(x) + \langle \lambda, Gx - h \rangle + \langle \nu, Ax - b \rangle \\ &= -\langle \lambda, h \rangle - \langle \nu, b \rangle - \sup_x \left(\langle -G^\top \lambda - A^\top \nu, x \rangle - f(x) \right) \\ &= -\langle \lambda, h \rangle - \langle \nu, b \rangle - f^*(-G^\top \lambda - A^\top \nu) \end{aligned}$$

Example: Linear Programming

- $f(x) = \langle c, x \rangle$
- $f^*(y) = \begin{cases} 0 & y = c \\ \infty & \text{otherwise} \end{cases}$

$$\begin{aligned} \max \quad & -\langle \lambda, h \rangle - \langle \nu, b \rangle \\ \text{s.t.} \quad & -G^\top \lambda - A^\top \nu = c \\ & \lambda \geq 0 \end{aligned}$$

Example: Min Norm Solution

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & \|x\|_2 \\ \text{s.t.} & Ax = b \end{array}$$

$$\begin{array}{ll} \max\limits_{\nu} & \langle -b, \nu \rangle \\ \text{s.t.} & \|A^\top \nu\|_2 \leq 1 \end{array}$$

- $f(x) = \|x\|_2$
- $f^*(y) = \sup\limits_x (\langle y, x \rangle - \|x\|_2) = \begin{cases} 0 & \|y\|_2 \leq 1 \\ \infty & \|y\|_2 > 1 \end{cases}$

If $\|y\|_2 \leq 1$, then $\langle y, x \rangle \leq \|x\|_2$
If $\|y\|_2 > 1$, take $x = ty, t \to \infty$

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & Ax = b \\ & Gx \leq h \end{array}$$

$$\begin{array}{ll} \max & -\langle \lambda, h \rangle - \langle \nu, b \rangle - f^*(-G^\top \lambda - A^\top \nu) \\ \text{s.t.} & \lambda \geq 0 \end{array}$$

Example: Min Norm Solution

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & \|x\|_2 \\ \text{s.t.} & Ax = b \end{array}$$

$$\begin{array}{ll} \max\limits_{\nu} & \langle -b, \nu \rangle \\ \text{s.t.} & \|A^\top \nu\|_2 \leq 1 \end{array}$$

- $f(x) = \|x\|_2$
- $f^*(y) = \sup\limits_x (\langle y, x \rangle - \|x\|_2) = \begin{cases} 0 & \|y\|_2 \leq 1 \\ \infty & \|y\|_2 > 1 \end{cases}$

If $\|y\|_2 \leq 1$, then $\langle y, x \rangle \leq \|x\|_2$
If $\|y\|_2 > 1$, take $x = ty, t \to \infty$

- For a general norm, $f(x) = \|x\|$: $f^*(x) = \begin{cases} 0 & \|y\|_* \leq 1 \\ \infty & \|y\|_* > 1 \end{cases}$
- Dual norm: $\|y\|_* = \sup\limits_{\|x\| \leq 1} \langle y, x \rangle$

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & \|x\| \\ \text{s.t.} & Ax = b \end{array}$$

$$\begin{array}{ll} \max\limits_{\nu} & \langle -b, \nu \rangle \\ \text{s.t.} & \|A^\top \nu\|_* \leq 1 \end{array}$$

Combining Fenchel Conjugates

$$\boldsymbol{f}(x_1, x_2, \dots, x_n) = \sum_{i=1}^{n} f_i(x_i)$$

$$\boldsymbol{f}^*(\boldsymbol{y}) = \sup_x \big(\langle y, x\rangle - f(x)\big) = \sup_x \textstyle\sum_{i=1}^{n} \big(y_i x_i - f_i(x_i)\big)$$

$$= \textstyle\sum_{i=1}^{n} \sup_{x_i} \big(y_i x_i - f_i(x_i)\big) \;\; = \textstyle\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{f_i^*}(y_i)$$

Example: Linear Classification

$$\min_{w \in \mathbb{R}^n} \quad \sum_{i=1}^{m} \phi(y_i \langle w, x_i \rangle)$$

$$\max g$$

$$g = \inf_{w} \sum_{i=1}^{m} \phi(y_i \langle w, x_i \rangle)$$

Example: Linear Classification

$$\min_{w \in \mathbb{R}^n} \quad \sum_{i=1}^m \phi(y_i \langle w, x_i \rangle)$$

$$\begin{aligned} \min_{w \in \mathbb{R}^n, z \in \mathbb{R}^m} \quad & \sum_{i=1}^m \phi(z_i) \\ \text{s.t.} \quad & z_i = y_i \langle w, x_i \rangle \end{aligned}$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\sum_{i=1}^m \phi^*(\nu_i) \\ \text{s.t.} \quad & \sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

- $f(z, w) = \sum_{i=1}^m \phi(z_i) + \langle 0, w \rangle$
- $f^*(\zeta, \omega) = \sum_{i=1}^m \phi^*(\zeta_i) + \begin{cases} 0 & \omega = 0 \\ \infty & \omega \neq 0 \end{cases}$ $\qquad \zeta \in (\mathbb{R}^m)^*, \omega \in (\mathbb{R}^n)^*$

$\tilde{x}_i = y_i x_i$, i.e. $\tilde{X} = X \cdot \mathrm{diag}(y) \in \mathbb{R}^{n \times m}$

- Constraint: $[-I, \tilde{X}^\top][z; w] = 0$
- Dual objective: $-\langle \nu, 0 \rangle - f^*([I\nu, -\tilde{X}\nu]) = -\sum_{i=1}^m \phi^*(\nu_i)$ s.t. $\sum_{i=1}^m y_i \nu_i x_i = 0$

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & Ax = b, Gx \leq h \end{aligned}$$

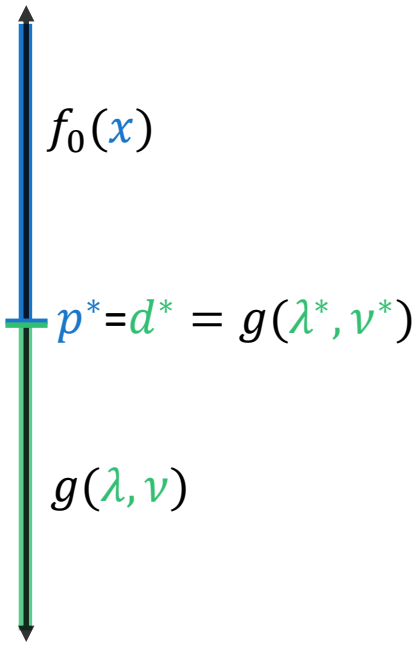
$$\begin{aligned} \max \quad & -\langle \lambda, h \rangle - \langle \nu, b \rangle - f^*(-G^\top \lambda - A^\top \nu) \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned}$$

Example: Linear Classification

$$\min_{w \in \mathbb{R}^n} \quad \sum_{i=1}^m \phi(y_i \langle w, x_i \rangle)$$

$$\min_{w \in \mathbb{R}^n, z \in \mathbb{R}^m} \quad \sum_{i=1}^m \phi(z_i)$$
$$\text{s.t.} \quad z_i = y_i \langle w, x_i \rangle$$

$$\max_{v \in \mathbb{R}^m} \quad -\sum_{i=1}^m \phi^*(v_i)$$
$$\text{s.t.} \quad \sum_{i=1}^m y_i v_i x_i = 0$$



Example: Linear Classification

$$\min_{w \in \mathbb{R}^n} \quad \sum_{i=1}^m \phi(y_i \langle w, x_i \rangle)$$

$$\begin{aligned} \min_{w \in \mathbb{R}^n, z \in \mathbb{R}^m} \quad & \sum_{i=1}^m \phi(z_i) \\ \text{s.t.} \quad & z_i = y_i \langle w, x_i \rangle \end{aligned}$$

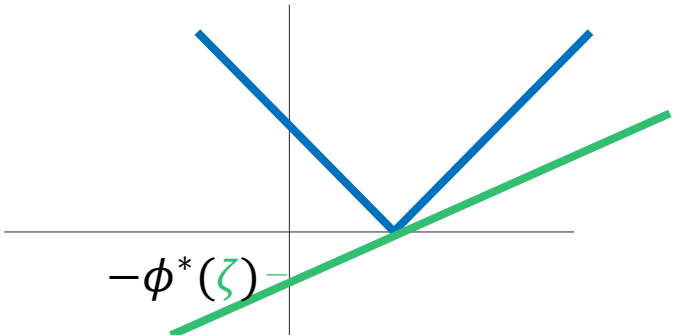
$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\sum_{i=1}^m \phi^*(\nu_i) \\ \text{s.t.} \quad & \sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

Absolute Error Loss:

$$\phi(z) = |z - 1| = |y_i \langle w, x_i \rangle - 1| = |\langle w, x_i \rangle - y_i|$$

$$\phi^*(\zeta) = \begin{cases} \zeta & -1 \le \zeta \le 1 \\ \infty & \text{otherwise} \end{cases}$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\sum_{i=1}^m \nu_i \\ \text{s.t.} \quad & -1 \le \nu_i \le 1 \\ & \sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$



Example: Linear Classification

$$\min_{w \in \mathbb{R}^n} \quad \sum_{i=1}^m \phi(y_i \langle w, x_i \rangle)$$

$$\begin{aligned} \min_{w \in \mathbb{R}^n, z \in \mathbb{R}^m} \quad & \sum_{i=1}^m \phi(z_i) \\ \text{s.t.} \quad & z_i = y_i \langle w, x_i \rangle \end{aligned}$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\sum_{i=1}^m \phi^*(\nu_i) \\ \text{s.t.} \quad & \sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

Absolute Error Loss:

$$\phi(z) = |z-1| = |y_i \langle w, x_i \rangle - 1| = |\langle w, x_i \rangle - y_i|$$

$$\phi^*(\zeta) = \begin{cases} \zeta & -1 \le \zeta \le 1 \\ \infty & \text{otherwise} \end{cases}$$

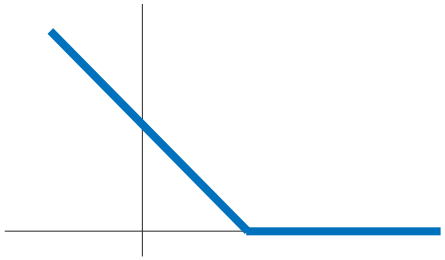
$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\sum_{i=1}^m \nu_i \\ \text{s.t.} \quad & -1 \le \nu_i \le 1 \\ & \sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

Squared Loss:

$$\phi(z-1) = \tfrac{1}{2}(z-1)^2 = (\langle w, x_i \rangle - y_i)^2$$

$$\phi^*(\zeta) = \tfrac{1}{2}\zeta^2 + \zeta$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\tfrac{1}{2}\sum_{i=1}^m \left(\nu_i + \nu_i^2\right) \\ \text{s.t.} \quad & \sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$



Hinge Loss:

$$\phi(z) = [1 - z]_+$$

$$\phi^*(y) = \begin{cases} y & -1 \le y \le 0 \\ \infty & \text{otherwise} \end{cases}$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\sum_{i=1}^m \nu_i \\ \text{s.t.} \quad & -1 \le \nu_i \le 0 \\ & \sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

Absolute Error Loss:

$$\phi(z) = |z - 1| = |y_i \langle w, x_i \rangle - 1| = |\langle w, x_i \rangle - y_i|$$

$$\phi^*(\zeta) = \begin{cases} \zeta & -1 \le \zeta \le 1 \\ \infty & \text{otherwise} \end{cases}$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\sum_{i=1}^m \nu_i \\ \text{s.t.} \quad & -1 \le \nu_i \le 1 \\ & \sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

Squared Loss:

$$\phi(z) = \tfrac{1}{2} z^2 = (\langle w, x_i \rangle - y_i)^2$$

$$\phi^*(\zeta) = \tfrac{1}{2} \zeta^2$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\tfrac{1}{2} \sum_{i=1}^m \nu_i^2 \\ \text{s.t.} \quad & \sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

Logistic Regression:

$$\phi(z) = \log(1 + e^{-z})$$

$$h(p) = -p \log p - (1-p)\log(1-p)$$

$$\phi^*(\zeta) = \begin{cases} -h(-\zeta) & -1 \le \zeta \le 0 \\ \infty & \text{otherwise} \end{cases}$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & \textstyle\sum_{i=1}^m h(-\nu_i) \\ \text{s.t.} \quad & -1 \le \nu_i \le 0 \\ & \textstyle\sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

Hinge Loss:

$$\phi(z) = [1 - z]_+$$

$$\phi^*(\zeta) = \begin{cases} \zeta & -1 \le \zeta \le 0 \\ \infty & \text{otherwise} \end{cases}$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\textstyle\sum_{i=1}^m \nu_i \\ \text{s.t.} \quad & -1 \le \nu_i \le 0 \\ & \textstyle\sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

Absolute Error Loss:

$$\phi(z) = |z - 1| = |y_i \langle w, x_i \rangle - 1| = |\langle w, x_i \rangle - y_i|$$

$$\phi^*(\zeta) = \begin{cases} \zeta & -1 \le \zeta \le 1 \\ \infty & \text{otherwise} \end{cases}$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\textstyle\sum_{i=1}^m \nu_i \\ \text{s.t.} \quad & -1 \le \nu_i \le 1 \\ & \textstyle\sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

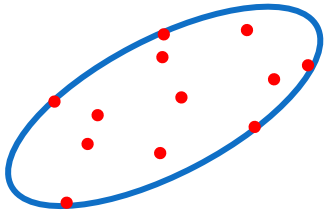
Squared Loss:

$$\phi(z) = \tfrac{1}{2} z^2 = (\langle w, x_i \rangle - y_i)^2$$

$$\phi^*(\zeta) = \tfrac{1}{2} \zeta^2$$

$$\begin{aligned} \max_{\nu \in \mathbb{R}^m} \quad & -\tfrac{1}{2} \textstyle\sum_{i=1}^m \nu_i^2 \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^m y_i \nu_i x_i = 0 \end{aligned}$$

Minimum Volume Covering Ellipsoid



- Problem: Given points $\{a_i \in \mathbb{R}^n\}$ find minimum volume ellipsoid centered at origin containing all points

$$\mathcal{E} = \{z \mid \|Pz\|^2 \leq 1\} = \{z \mid z^\top X z \leq 1\}$$

$X = P^\top P \succcurlyeq 0$

$$Volume(\mathcal{E}) \propto \sqrt{\det X^{-1}}$$

$S^n \subset \mathbb{R}^{n \times n}$ Symmetric

$$\begin{aligned} \min_{X \in S^n} \quad & -\log\det X \\ \text{s.t.} \quad & a_i^\top X a_i \leq 1 \quad i = 1..m \end{aligned}$$

$$a_i^\top X a_i = \sum_{jk} (a_i[j] a_i[k]) X[j,k] = \langle a_i a_i^\top, X \rangle$$

$$\log\det X = \log \prod_i \lambda_i(X) \overset{\text{def}}{=} \sum_i \log \lambda_i(X) = \begin{cases} \log(\det X) & \text{if } X \succcurlyeq 0 \\ \infty & \text{otherwise} \end{cases}$$

Dual of MVCE

$$\begin{aligned} \min_{X \in S^n} \quad & -\log\det X \\ \text{s.t.} \quad & a_i^\top X a_i \le 1 \quad i = 1..m \\ & X \succ 0 \end{aligned}$$

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m} \quad & -\sum \lambda_i + n + \log\det\left(\textstyle\sum_i \lambda_i a_i a_i^\top\right) \\ \text{s.t.} \quad & \lambda_i \ge 0 \\ & \textstyle\sum_i \lambda_i a_i a_i^\top \succ 0 \end{aligned}$$

- Fenchel conjugate of $f(X) = -\log\det X$:

$Y = \nabla f(X) = -X^{-1}$ ➔ $X = -Y^{-1}$ if $Y \prec 0$

$\langle A, B\rangle = \sum_{jk} A[j,k]B[j,k] = tr(A^\top B)$

$$f^*(Y) = \langle Y, -Y^{-1}\rangle + \log\det(-Y^{-1}) = -tr(I) - \log\det(-Y)$$

And so:
$$f^*(Y) = \begin{cases} -n - \log\det(-Y) & \text{if } Y \prec 0 \\ \infty & \text{otherwise} \end{cases}$$

- Strong Duality?
  - Yes! Large enough ellipsoid always feasible, and $\log\det X$ finite.

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & Ax = b,\ Gx \le h \end{aligned}$$

$$\begin{aligned} \max \quad & -\langle \lambda, h\rangle - \langle \nu, b\rangle - f^*(-G^\top \lambda - A^\top \nu) \\ \text{s.t.} \quad & \lambda \ge 0 \end{aligned}$$

Matrix Inequalities (Semi Definite Constraints)

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f_0(x) \\ s.t. \quad & f_i(x) \preccurlyeq 0 \\ & h_i(x) = 0 \end{aligned}$$

$f_0 \colon \mathbb{R}^n \to \mathbb{R}$ convex

$f_i \colon \mathbb{R}^n \to S^{k_i} \subset \mathbb{R}^{k_i \times k_i}$ **matrix-convex**:

$$f_i(\theta x + (1-\theta)x') \preccurlyeq \theta f_i(x) + (1-\theta) f_i(x')$$

$h_i \colon \mathbb{R}^n \to \mathbb{R}$ linear (or $\to S^k$, doesn't matter much)

Sometimes convenient to represent $x \in S^n$: doesn't matter much, its just a vector space of dimensionality $n(n+1)/2$

Semi-Definite Programming (SDP): f_0, f_i **linear**