8. Robust models

- stochastic programming
- mean field approximation
- finite event set and sample average approximation
- worst-case optimization
- worst-case robust approximation
- robust linear discrimination
- geometric interpretation and Lagrange duality
- support vector classifiers

Stochastic programming

a **stochastic program** has the form

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\, f_0(x,\omega) \\ \text{subject to} & \mathbf{E}\, f_i(x,\omega) \leq 0, \quad i=1,\ldots,m \end{array}$$

- variable $x \in \mathbf{R}^n$; parameter $\omega \in \mathbf{R}^q$ is random variable
- expectation is taken over ω
- constraints called **stochastic constraints**

convex stochastic program if f_0 and f_i are convex in x for each ω

- a few can be solved efficiently
- in general, evaluating the expectation is difficult

Stochactic least squares

$$\text{minimize} \quad \mathbf{E}\,\|Ax - b\|_2^2$$

- problem data $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$ are random variables
- objective expanded as

$$\begin{aligned}
\mathbf{E}\,\|Ax - b\|_2^2 &= \mathbf{E}\,(Ax - b)^T(Ax - b) \\
&= \mathbf{E}(x^T A^T Ax - 2b^T Ax + b^T b) \\
&= x^T(\mathbf{E}\,A^T A)x - 2(\mathbf{E}\,b^T A)x + \mathbf{E}\,b^T b
\end{aligned}$$

so equivalent to a convex QP if second moments of A and b are known

example: $A = \bar{A} + U$ with U zero mean random, $\bar{A}$, b deterministic

$$\text{minimize} \quad \mathbf{E}\,\|Ax - b\|_2^2 = \|\bar{A}x - b\|_2^2 + x^T(\mathbf{E}\,U^T U)x$$

let $P = \mathbf{E}\, U^T U$ be the covariance of U, then problem equivalent to

$$\text{minimize} \quad \|\bar{A}x - b\|_2^2 + \|P^{1/2}x\|_2^2$$

has analytical solution $x^\star = (\bar{A}^T\bar{A} + P)^{-1}\bar{A}^T b$

- for $U \in \mathbf{R}^{m\times n}$ with IID $\mathcal{N}(0, \sigma^2)$ entries, $P = m\sigma^2 I$

$$x^\star = (\bar{A}^T\bar{A} + m\sigma^2 I)^{-1}\bar{A}^T b$$

  equivalent to Tikhonov regularization with weight $\gamma = m\sigma^2$

Mean field approximation

$$\begin{array}{ll} \text{minimize} & f_0(x, \mathbf{E}\,\omega) \\ \text{subject to} & f_i(x, \mathbf{E}\,\omega) \leq 0, \quad i = 1, \ldots, m \end{array}$$

- ignore the randomness in ω and replace it with its mean
- sometimes called *certainty equivalent problem*
- convex if f_i convex in x for each ω

relaxation property: if f_i convex in ω for each x, by Jensen's inequality

$$f_i(x, \mathbf{E}\,\omega) \leq \mathbf{E}\, f_i(x, \omega)$$

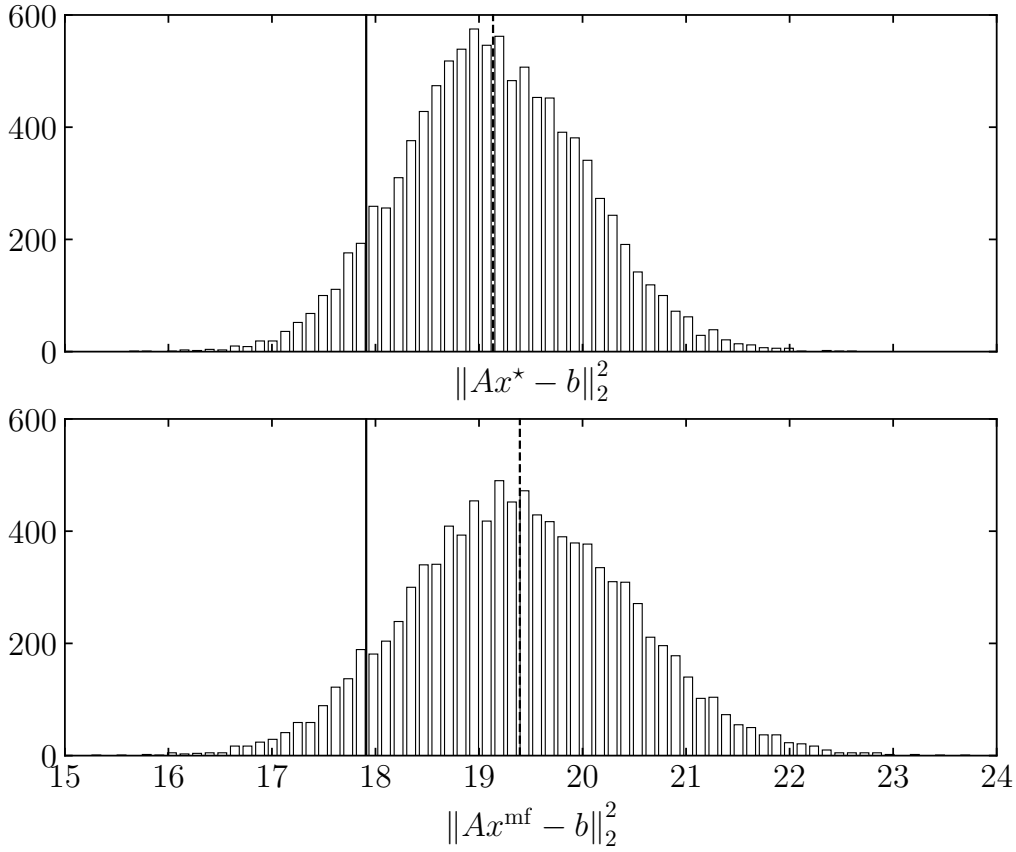
- objective lower bound: $f_0(x, \mathbf{E}\,\omega) \leq \mathbf{E}\, f_0(x, \omega)$
- constraint relaxation: $\mathbf{E}\, f_i(x, \omega) \leq 0 \implies f_i(x, \mathbf{E}\,\omega) \leq 0$, $i = 1, \ldots, m$

example: stochastic least squares

$$\text{minimize} \quad \mathbf{E}\,\|Ax-b\|_2^2 \;, \quad A=\bar{A}+U, \quad U\sim\mathcal{N}(0,m\sigma^2 I)$$

mean field approximation minimizes $\|\bar{A}x-b\|_2^2$

- $x^\star$: stochastic solution
- x^{mf}: mean field solution
- solid line shows mean field objective lower bound



Finite event set

for random parameter $\omega \in \{\omega_1, \ldots, \omega_k\} \subseteq \mathbf{R}^q$ with

$$\mathbf{prob}(\omega = \omega_i) = p_i \quad (p \succeq 0,\ \mathbf{1}^T p = 1)$$

stochastic program reduces to deterministic program

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^k p_i f_0(x, \omega_i) \\ \text{subject to} & \sum_{i=1}^k p_i f_j(x, \omega_i) \leq 0, \quad j = 1, \ldots, m \end{array}$$

example: stochastic linear approximation

$$\text{minimize} \quad \mathbf{E}\, \|Ax - b\|$$

if $A \in \{A_1, \ldots, A_k\}$ with $\mathbf{prob}(A = A_i) = p_i$, reduces to

$$\text{minimize} \quad \sum_{i=1}^k p_i \|A_i x - b\|$$

- for ℓ_1-norm, reduces to LP

$$\begin{array}{ll} \text{minimize (over } x,\, t) & p^T t \\ \text{subject to} & \|A_i x - b\|_1 \leq t_i, \quad i = 1, \ldots, k \end{array}$$

$$\iff \quad \begin{array}{ll} \text{minimize (over } x,\, u_i) & \sum_{i=1}^k p_i \mathbf{1}^T u_i \\ \text{subject to} & -u_i \preceq A_i x - b \preceq u_i, \quad i = 1, \ldots, k \end{array}$$

- for ℓ_∞-norm, reduces to LP

$$\begin{array}{ll} \text{minimize (over } x,\, t) & p^T t \\ \text{subject to} & -t_i \mathbf{1} \preceq A_i x - b \preceq t_i \mathbf{1}, \quad i = 1, \ldots, k \end{array}$$

Sample average approximation

approximate expectation by finite sample average

$$\begin{array}{ll} \text{minimize} & (1/N)\sum_{i=1}^{N} f_0(x,\omega_i) \\ \text{subject to} & (1/N)\sum_{i=1}^{N} f_j(x,\omega_i) \leq 0, \quad j=1,\ldots,m \end{array}$$

- $\omega_1,\ldots,\omega_N$ are IID samples of ω
- solution x^{saa} converges to stochastic solution $x^\star$ as $N \to \infty$
- optimal value p^{saa} lower bounds $p^\star$, tight as $N \to \infty$

example: stochastic ℓ_1-norm approximation

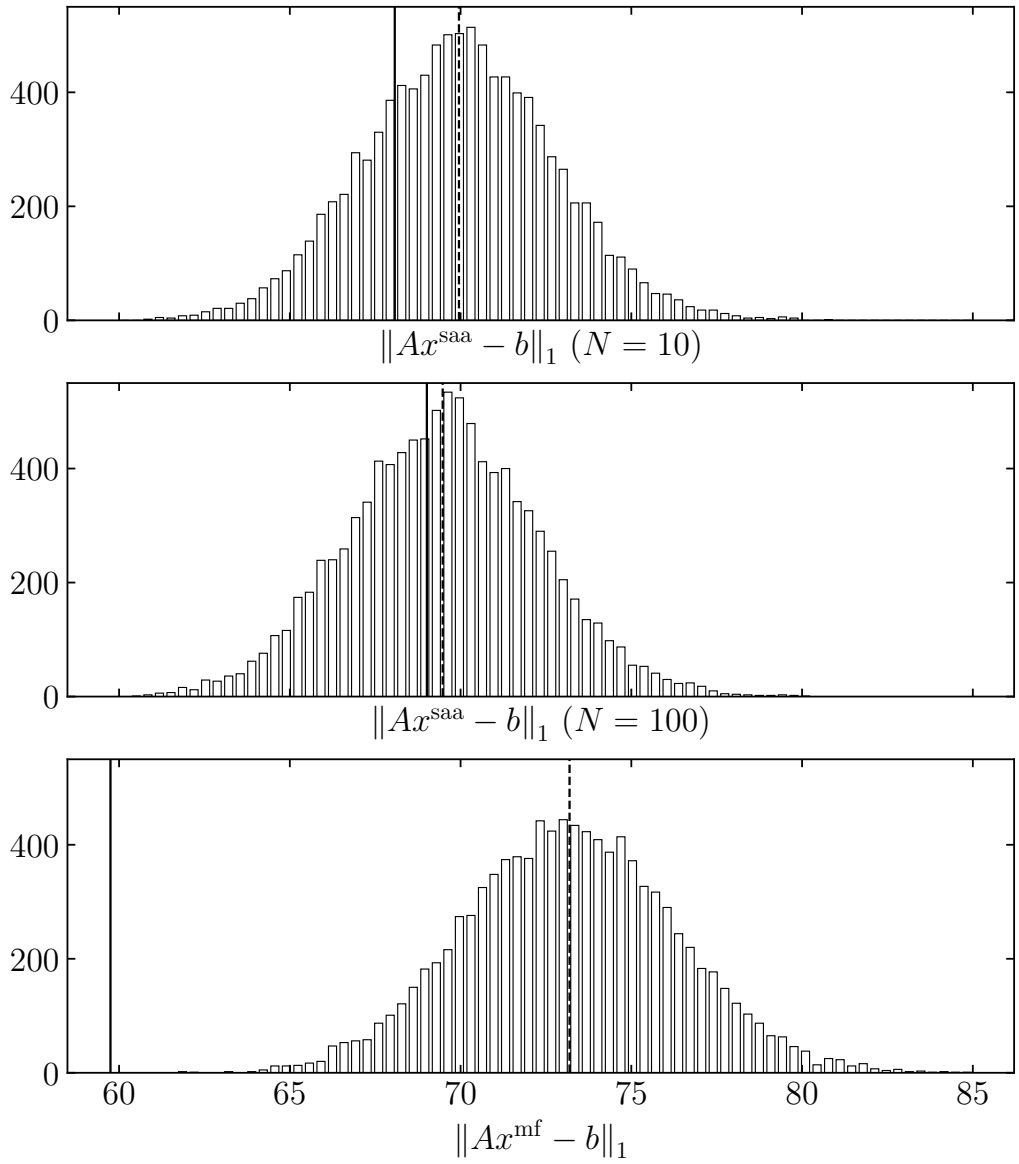
$$\text{minimize} \quad \mathbf{E}\,\|Ax-b\|_1$$

for $A = \bar{A} + U$, $b = \bar{b} + v$ with U, v zero mean IID Gaussian

sample average approximation:

$$\text{minimize} \quad (1/N)\sum_{i=1}^{N} \|A_i x - b_i\|_1$$

optimal value shown solid; mean field solution x^{mf} shown bottom:



Worst-case optimization

$$\begin{array}{ll} \text{minimize} & \sup_{\omega \in U} f_0(x, \omega) \\ \text{subject to} & \sup_{\omega \in U} f_i(x, \omega) \leq 0, \quad i = 1, \ldots, m \end{array}$$

- variable $x \in \mathbf{R}^n$
- $f_i : \mathbf{R}^n \times \mathbf{R}^q \to \mathbf{R}$, parameterized by $\omega \in U \subseteq \mathbf{R}^q$
- U sometimes called the *uncertainty set*

interpretation: choose x so that for *all* $\omega \in U$

- $f_0(x, \omega)$ is minimized
- $f_i(x, \omega) \leq 0$ for $i = 1, \ldots, m$ are satisfied

convex worst-case optimization if f_i convex in x for each $\omega \in U$

- evaluating the supremum is often difficult; with a few exceptions

example: for finite uncertainty set $U = \{\omega_1, \ldots, \omega_k\}$,

$$\begin{array}{lll} & \text{minimize} & \max_{i=1,\ldots,k} f_0(x, \omega_i) \\ & \text{subject to} & \max_{i=1,\ldots,k} f_j(x, \omega_i) \leq 0, \quad j = 1, \ldots, m \end{array}$$

$$\begin{array}{lll} & \text{minimize} & t \\ \iff & \text{subject to} & f_0(x, \omega_i) \leq t, \quad i = 1, \ldots, k \\ & & f_j(x, \omega_i) \leq 0, \quad i = 1, \ldots, k, \ j = 1, \ldots, m \end{array}$$

- convex program (in x and t) if f_i convex in x for each $\omega \in U$

Worst-case robust approximation

$$\text{minimize} \quad \sup\{\|Ax - b\|_2^2 \mid A \in \mathcal{A}\}$$

- variable $x \in \mathbf{R}^n$
- $A \in \mathcal{A} \subseteq \mathbf{R}^{m \times n}$ is uncertain; assume $b \in \mathbf{R}^m$ fixed

example: for A on a line segment between $A_0 - A_1$ and $A_0 + A_1$, *i.e.*,

$$\mathcal{A} = \{A_0 + uA_1 \mid |u| \leq 1\}$$

- nominal solution ignores uncertainty in A, *i.e.*,

$$\text{minimize} \quad \|A_0 x - b\|_2^2$$

- stochastic solution minimizes expected objective, *i.e.*,

$$\text{minimize} \quad \mathbf{E}\,\|(A_0 + uA_1)x - b\|_2^2 \;, \quad u \sim \mathcal{U}(-1, 1)$$

which is equivalent to

$$\text{minimize} \quad \|A_0 x - b\|_2^2 + \|P^{1/2} x\|_2^2,$$

where $P = \mathbf{E}\,(uA_1)^T(uA_1) = \mathbf{E}\,u^2 A_1^T A_1 = (1/3)A_1^T A_1$

- worst-case objective is

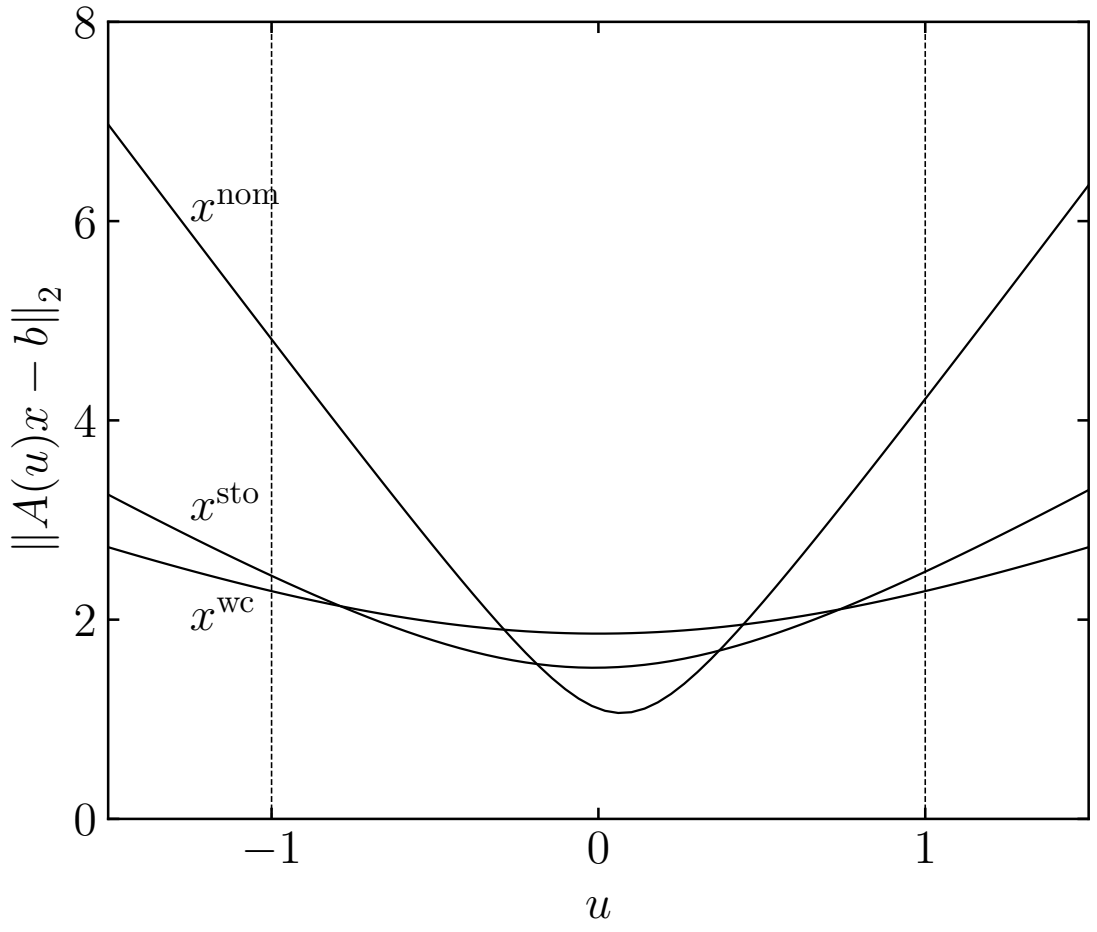
$$\begin{aligned} \sup_{A \in \mathcal{A}} \|Ax - b\|_2^2 &= \sup_{|u| \leq 1} \|(A_0 + uA_1)x - b\|_2^2 \\ &= \max\{\|(A_0 - A_1)x - b\|_2^2, \|(A_0 + A_1)x - b\|_2^2\} \end{aligned}$$

(since $\|(A_0 + uA_1)x - b\|_2^2$ is convex in u), so we have the problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & \|(A_0 - A_1)x - b\|_2^2 \leq t \\ & \|(A_0 + A_1)x - b\|_2^2 \leq t \end{array}$$

x^{nom}, x^{sto}, x^{wc}: nominal, stochastic, and worst-case solutions

evaluate $\|A(u)x - b\|_2$ for $u \in [-1, 1]$:



Robust linear discrimination

linear discrimination problem: given two groups of points $\{x_1, \ldots, x_M\}$ and $\{y_1, \ldots, y_N\}$ in $\mathbf{R}^n$,

$$\begin{array}{ll} \text{find} & (a, b) \\ \text{subject to} & a^T x_i - b \geq 1, \quad i = 1, \ldots, M \\ & a^T y_i - b \leq -1, \quad i = 1, \ldots, N \end{array}$$

- solution set is polyhedron

$$\left\{ z \in \mathbf{R}^{n+1} \;\middle|\; \begin{array}{l} z^T \tilde{x}_i \geq 1, \quad i = 1, \ldots, M \\ z^T \tilde{y}_i \leq -1, \quad i = 1, \ldots, N \end{array} \right\}$$

where

$$z = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \tilde{x}_i = \begin{bmatrix} x_i \\ -1 \end{bmatrix}, \quad \tilde{y}_i = \begin{bmatrix} y_i \\ -1 \end{bmatrix}$$

robust linear discrimination:

$$\begin{array}{ll} \text{minimize} & \|a\|_2 \\ \text{subject to} & a^T x_i - b \geq 1, \quad i = 1, \ldots, M \\ & a^T y_i - b \leq -1, \quad i = 1, \ldots, N \end{array}$$

- find normal vector a of separating hyperplane with least norm

Geometric interpretation

change of variables $a = \tilde{a}/t$, $b = \tilde{b}/t$ with $t > 0$ gives

$$\begin{array}{ll} \text{minimize} & \|\tilde{a}\|_2/t \\ \text{subject to} & \tilde{a}^T x_i - \tilde{b} \geq t, \quad i = 1, \ldots, M \\ & \tilde{a}^T y_i - \tilde{b} \leq -t, \quad i = 1, \ldots, N \\ & t > 0 \end{array}$$

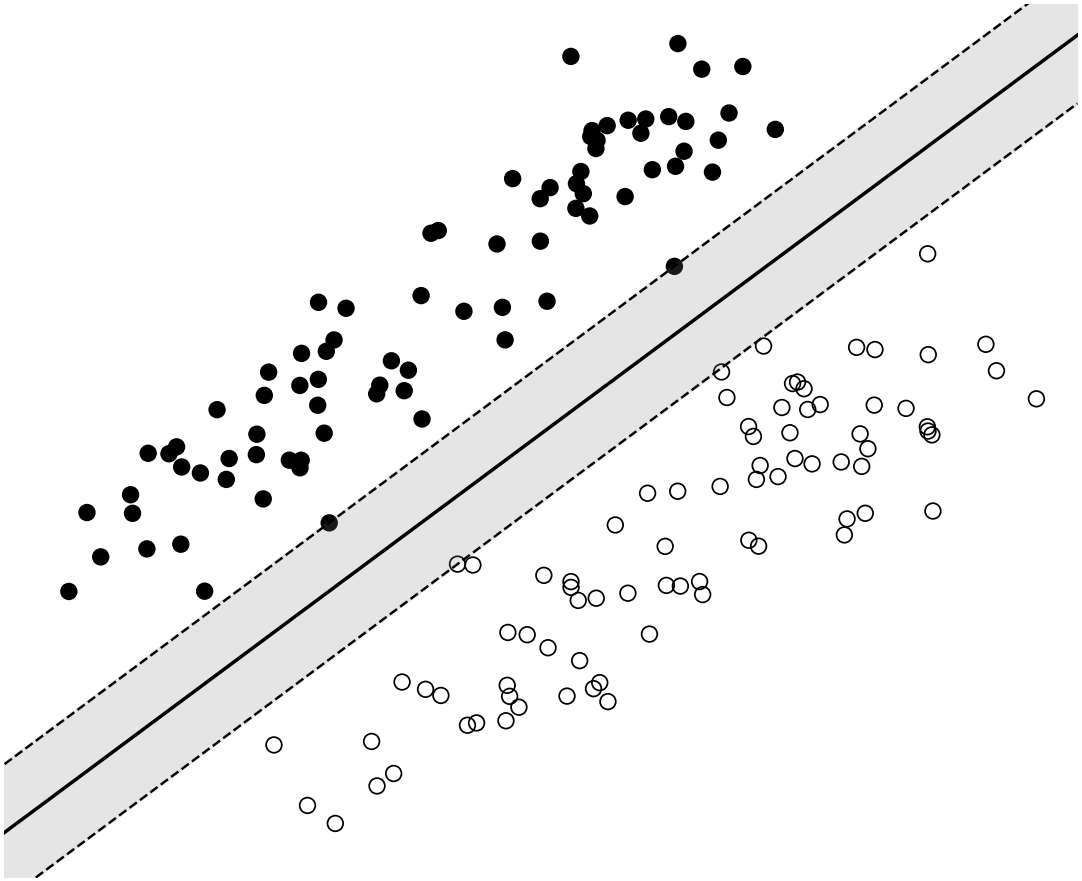
then to remove ambiguity, we bound $\|\tilde{a}\|_2 \leq 1$ and maximize t:

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & \tilde{a}^T x_i - \tilde{b} \geq t, \quad i = 1, \ldots, M \\ & \tilde{a}^T y_i - \tilde{b} \leq -t, \quad i = 1, \ldots, N \\ & t > 0, \quad \|\tilde{a}\|_2 \leq 1 \end{array}$$

interpretation: maximize the distance to the closest point in either group

for optimal point $(\tilde{a}^\star, \tilde{b}^\star, t^\star)$,

- $\|\tilde{a}^\star\|_2 = 1$ (by homogeneity of constraints)
- $2t^\star$ is the (thickest) separating slab width between two groups of points



Interpretation via Lagrange duality

primal problem of robust linear discrimination:

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & \tilde{a}^T x_i - \tilde{b} \geq t, \quad i = 1, \ldots, M \\ & \tilde{a}^T y_i - \tilde{b} \leq -t, \quad i = 1, \ldots, N \\ & \|\tilde{a}\|_2 \leq 1 \end{array}$$

- $t^\star > 0 \iff$ two groups of points are linearly separable

Lagrangian:

$$\begin{aligned} & L((\tilde{a}, \tilde{b}, t), (u, v, w)) \\ = \; & -t + \textstyle\sum_{i=1}^M u_i (t - \tilde{a}^T x_i + \tilde{b}) + \sum_{i=1}^N v_i (t + \tilde{a}^T y_i - \tilde{b}) + w(\|\tilde{a}\|_2 - 1) \\ = \; & (\mathbf{1}^T u + \mathbf{1}^T v - 1)t + (\mathbf{1}^T u - \mathbf{1}^T v)\tilde{b} \\ & -\tilde{a}^T \left(\textstyle\sum_{i=1}^M u_i x_i - \sum_{i=1}^N v_i y_i \right) + w(\|\tilde{a}\|_2 - 1) \end{aligned}$$

dual function $g(u,v,w) = \inf_{(\tilde{a},\tilde{b},t)} L((\tilde{a},\tilde{b},t),(u,v,w))$ bounded iff

$$\mathbf{1}^T u + \mathbf{1}^T v = 1, \quad \mathbf{1}^T u = \mathbf{1}^T v \quad \Longleftrightarrow \quad \mathbf{1}^T u = 1/2, \quad \mathbf{1}^T v = 1/2$$

in this case, we have

$$\begin{aligned} g(u,v,w) &= \inf_{\tilde{a}} \left(-\tilde{a}^T \left(\sum_{i=1}^M u_i x_i - \sum_{i=1}^N v_i y_i \right) + w(\|\tilde{a}\|_2 - 1) \right) \\ &= -w + \inf_{\tilde{a}} \left(-\tilde{a}^T \left(\sum_{i=1}^M u_i x_i - \sum_{i=1}^N v_i y_i \right) + w\|\tilde{a}\|_2 \right) \\ &\geq -w + \inf_{\tilde{a}} \left(\|\tilde{a}\|_2 \left(w - \left\| \sum_{i=1}^M u_i x_i - \sum_{i=1}^N v_i y_i \right\|_2 \right) \right) \end{aligned}$$

inequality follows from Cauchy-Schwarz, so

$$g(u,v,w) = \begin{cases} -w, & w \geq \left\| \sum_{i=1}^M u_i x_i - \sum_{i=1}^N v_i y_i \right\|_2 \\ -\infty, & \text{otherwise} \end{cases}$$

dual problem of robust linear discrimination:

$$\begin{array}{ll} \text{maximize} & -w \\ \text{subject to} & w \geq \left\| \sum_{i=1}^M u_i x_i - \sum_{i=1}^N v_i y_i \right\|_2 \\ & \mathbf{1}^T u = 1/2, \quad \mathbf{1}^T v = 1/2 \\ & u \succeq 0, \quad v \succeq 0 \end{array}$$

is the epigraph form of

$$\begin{array}{ll} \text{minimize} & \left\| \sum_{i=1}^M u_i x_i - \sum_{i=1}^N v_i y_i \right\|_2 \\ \text{subject to} & \mathbf{1}^T u = 1/2, \quad \mathbf{1}^T v = 1/2 \\ & u \succeq 0, \quad v \succeq 0 \end{array}$$

interpretation: find closest points in convex hulls of two groups, *i.e.*,

$$\begin{array}{ll} \text{minimize} & (1/2)\|x - y\|_2 \\ \text{subject to} & x \in \mathbf{conv}\{x_1, \ldots, x_M\} \\ & y \in \mathbf{conv}\{y_1, \ldots, y_N\} \end{array}$$

Support vector classifiers

for two groups of points $\{x_1, \ldots, x_M\}$ and $\{y_1, \ldots, y_N\}$ in $\mathbf{R}^n$ that are not linearly separable, consider relaxation

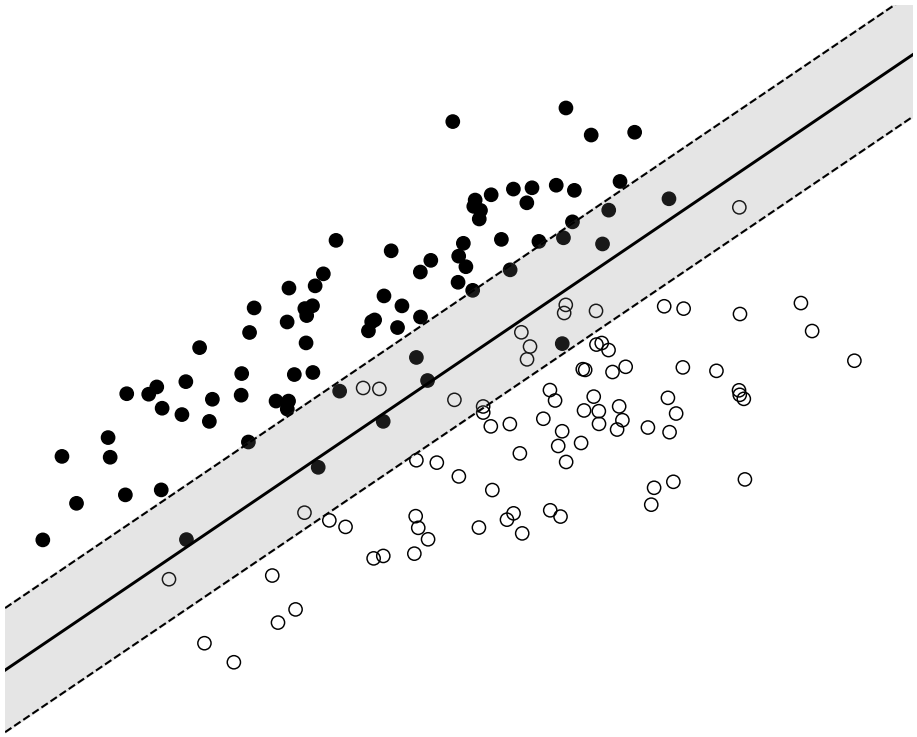
$$\begin{array}{ll} \text{find} & (a, b, u, v) \\ \text{subject to} & a^T x_i - b \geq 1 - u_i, \quad i = 1, \ldots, M \\ & a^T y_i - b \leq -1 + v_i, \quad i = 1, \ldots, N \\ & u \succeq 0, \quad v \succeq 0 \end{array}$$

- slack variables u, v represent margin violations for each point
- always feasible, but may have $u_i, v_i > 0$ for some i
  - within margin, but not misclassified if $0 < u_i, v_i < 1$
  - misclassification if $u_i, v_i \geq 1$
- linearly separable if exist optimal point with $u = v = 0$

Margin violation penalties

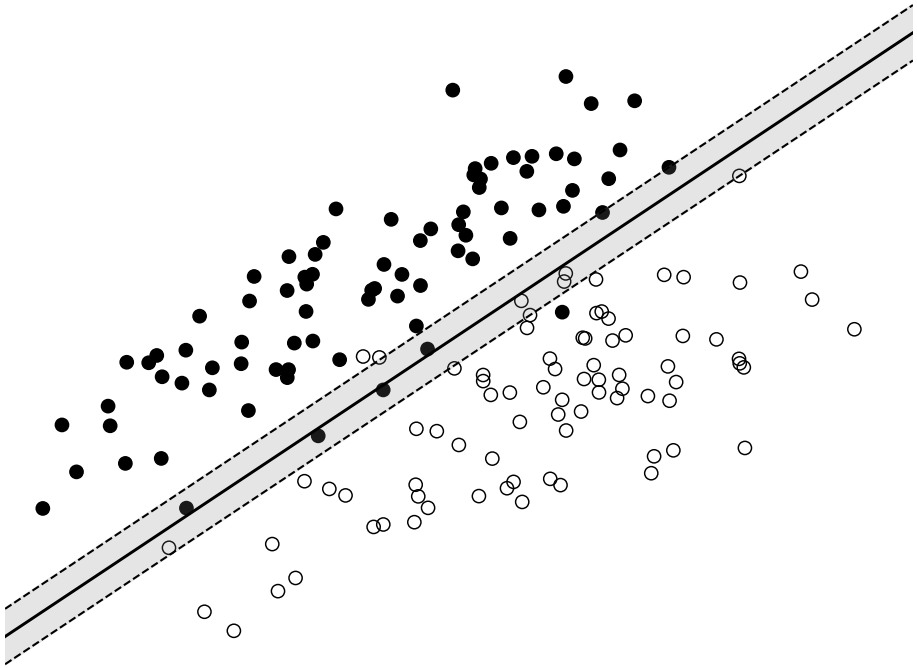
least squares penalty:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{M} u_i^2 + \sum_{i=1}^{N} v_i^2 \\ \text{subject to} & a^T x_i - b \geq 1 - u_i, \quad i = 1, \ldots, M \\ & a^T y_i - b \leq -1 + v_i, \quad i = 1, \ldots, N \\ & u \succeq 0, \quad v \succeq 0 \end{array}$$



sparse margin violations: heuristic for minimizing $\mathbf{card}\, u + \mathbf{card}\, v$,

$$
\begin{array}{ll}
\text{minimize} & \mathbf{1}^T u + \mathbf{1}^T v \\
\text{subject to} & a^T x_i - b \geq 1 - u_i, \quad i = 1, \ldots, M \\
& a^T y_i - b \leq -1 + v_i, \quad i = 1, \ldots, N \\
& u \succeq 0, \quad v \succeq 0
\end{array}
$$



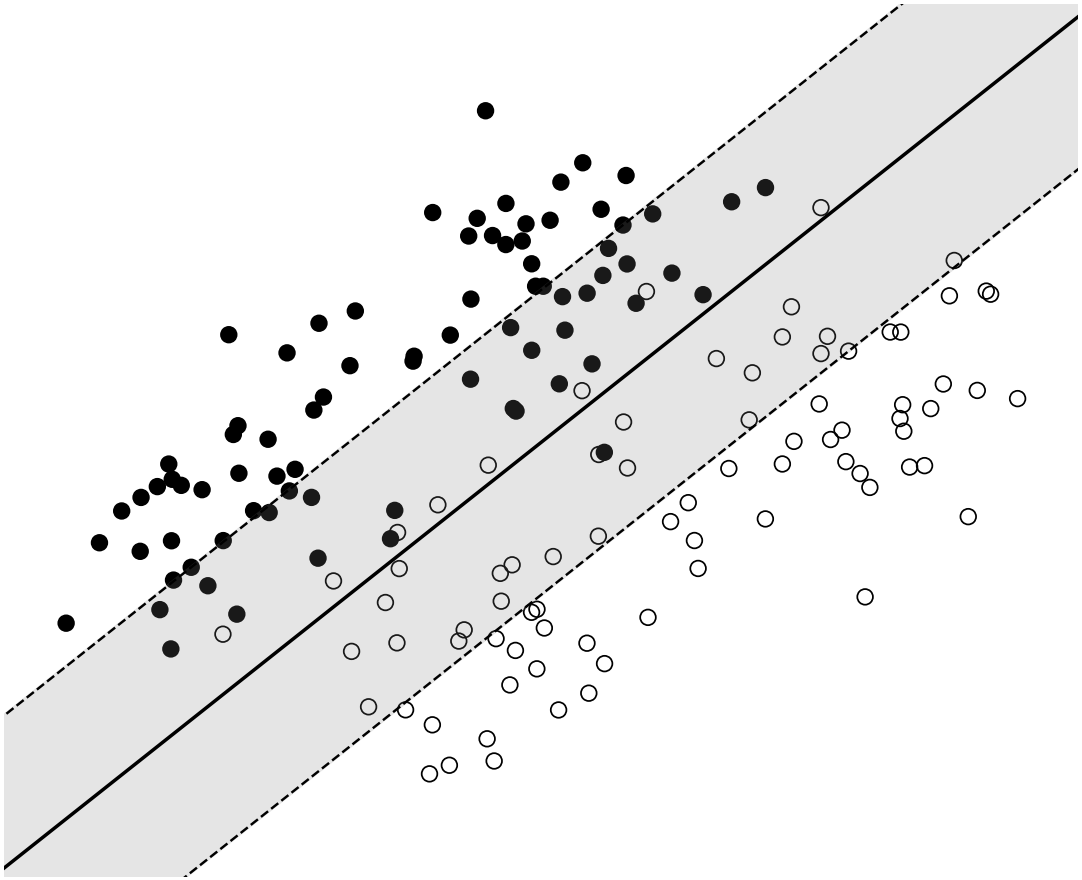
Standard form support vector classifier

trade-off between margin width and margin violation penalty:

$$\begin{array}{ll} \text{minimize} & \|a\|_2 + \gamma(\mathbf{1}^T u + \mathbf{1}^T v) \\ \text{subject to} & a^T x_i - b \geq 1 - u_i, \quad i = 1, \ldots, M \\ & a^T y_i - b \leq -1 + v_i, \quad i = 1, \ldots, N \\ & u \succeq 0, \quad v \succeq 0 \end{array}$$

- variables $a \in \mathbf{R}^n$, $b \in \mathbf{R}$, $u \in \mathbf{R}^M$, $v \in \mathbf{R}^N$
- $\gamma > 0$ is regularization weight
- separating slab width is $2/\|a\|_2$
- equals to feasibility problem as $\gamma \to \infty$ (no margin violations allowed)

for $\gamma = 0.01$:



for $\gamma = 1$:

