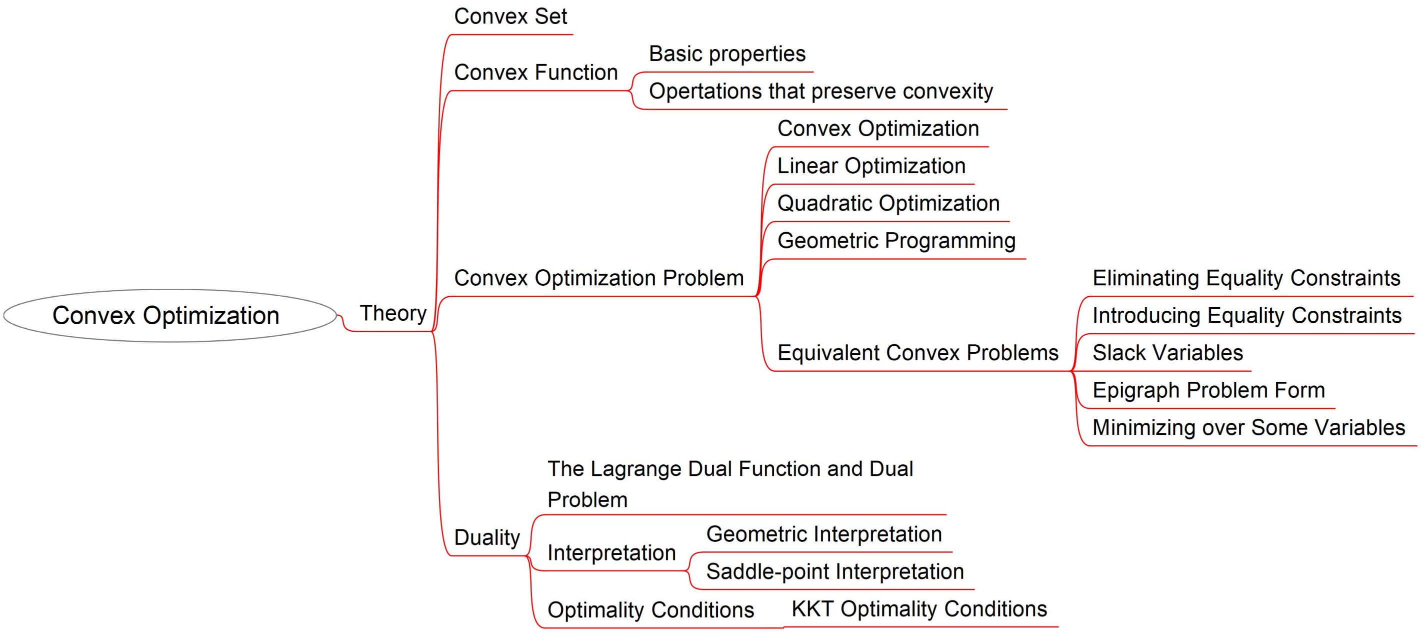
Convex Optimization

Zhenrong
March 13, 2021

1 Theory


Convex Optimization
Theory
Convex Set
Convex Function
Basic properties
Opertations that preserve convexity
Convex Optimization Problem
Convex Optimization
Linear Optimization
Quadratic Optimization
Geometric Programming
Equivalent Convex Problems
Eliminating Equality Constraints
Introducing Equality Constraints
Slack Variables
Epigraph Problem Form
Minimizing over Some Variables
Duality
The Lagrange Dual Function and Dual Problem
Interpretation
Geometric Interpretation
Saddle-point Interpretation
Optimality Conditions
KKT Optimality Conditions

1.1 Convex Set

1.1.1 Euclidean Balls and Ellipsoids

A (Euclidean) ball in $\mathbf{R}^n$ has the form

$$B\left(x_c, r\right) = \left\{x \mid \left\|x - x_c\right\|_2 \leq r\right\} = \left\{x \mid \left(x - x_c\right)^T \left(x - x_c\right) \leq r^2\right\},$$

where $r > 0$, and $\|\cdot\|_2$ denotes the Euclidean norm.

A related family of convex sets is the ellipsoids, which have the form

$$\mathcal{E} = \left\{x \mid \left(x - x_c\right)^T P^{-1} \left(x - x_c\right) \leq 1\right\},$$

where $P = P^T \succ 0$, i.e., P is symmetric and positive definite. The vector $x_c \in \mathbf{R}^n$ is the center of the ellipsoid. The matrix P determines how far the ellipsoid extends. Consider $u = P^{-\frac{1}{2}} \left(x - x_c\right)$, by variable substitution, we have $u^T u = 1$, and it is a Euclidean balls. An understanding of the geometry induced by SVD is that the semi-axes of the ellipsoids is given by the singular value of $P^{-\frac{1}{2}}$.

1.2 Strong Convexity and Smoothness

A function f is strongly convex with parameter $m > 0$ (written m-strongly convex) provided that

$$f(x) - \frac{m}{2}\|x\|_2^2$$

is a convex function. Roughly speaking, this means that f is "as least as convex" as a quadratic function. Here are some important facts about f, which are equivalent conditions of strong convexity,

$$\begin{array}{l} (\nabla f(x) - \nabla f(y))^T(x-y) \geq m\|x-y\|^2, \forall x, y, \\ \nabla^2 f(x) \succeq mI, \forall x, \\ f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2. \end{array} \tag{1}$$

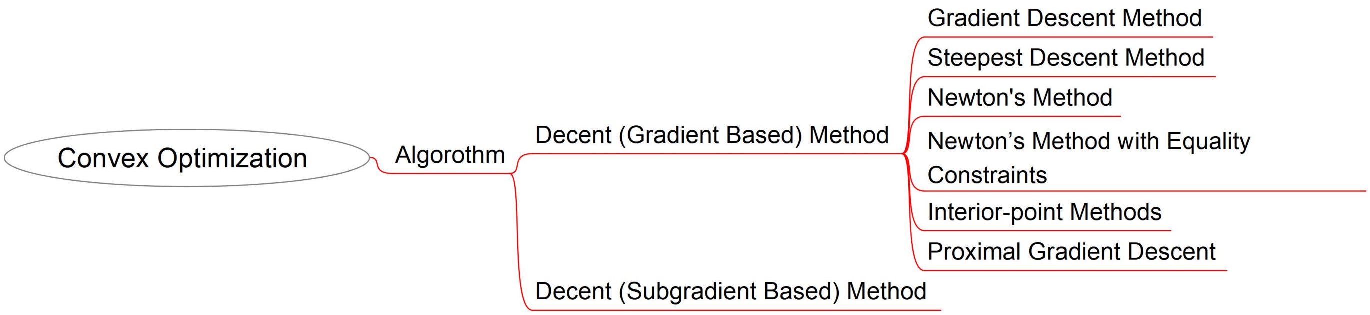
A function f is smooth with parameter L (written L-smooth) if $\nabla f(x)$ $\forall x, y, \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|, L > 0$.

The above function restrict the rate of changing of derivatives. Here are some important facts about f, which are equivalent conditions of $L-$ smooth.

$$\begin{array}{l} (\nabla f(x) - \nabla f(y))^T(x-y) \leq L\|x-y\|^2, \forall x, y, \\ \nabla^2 f(x) \preceq LI, \forall x, \\ f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2. \end{array} \tag{2}$$

The assumption of strong convexity and smoothness sometimes plays important roles in analysis of convergence.

2 Algorithm


Convex Optimization
Algorothm
Decent (Gradient Based) Method
Gradient Descent Method
Steepest Descent Method
Newton's Method
Newton's Method with Equality Constraints
Interior-point Methods
Proximal Gradient Descent
Decent (Subgradient Based) Method

2.1 Decent Method

The decent method produce a minimizing sequence $x^{(k)}, k = 1, \ldots,$ where

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$$

and $t^{(k)} > 0$ (except when $x^{(k)}$ is optimal). Δx is a vector in $\mathbf{R}^n$ called the step or search direction. The scalar $t^{(k)} \geq 0$ is called the step size or step length at iteration k.

The search direction in a descent method must satisfy

$$\nabla f \left(x^{(k)}\right)^T \Delta x^{(k)} < 0.$$

i.e., it must make an acute angle with the negative gradient. The process of a general descent method is as follows. It alternates between two steps: determining a descent direction Δx, and the selection of a step size t.

2.1.1 Line Search Method

Exact line search

One line search method sometimes used in practice is exact line search, in which t is chosen to minimize f along the ray $\{x + t\Delta x \mid t \geq 0\}$:

$$t = \operatorname{argmin}_{s \geq 0} f(x + s\Delta x).$$

An exact line search is used when the cost of the minimization problem with one variable, is low compared to the cost of computing the search direction itself. In some special cases the minimizer along the ray can be found analytically, and in others it can be computed efficiently. For example, if the objective function is quadratic, then the step length of exact line search can be given analytically.

Backtracking line search

Most line searches used in practice are inexact: the step length is chosen to approximately minimize f along the ray $\{x+t\Delta x \mid t \geq 0\}$. One inexact line search method is called backtracking line search.

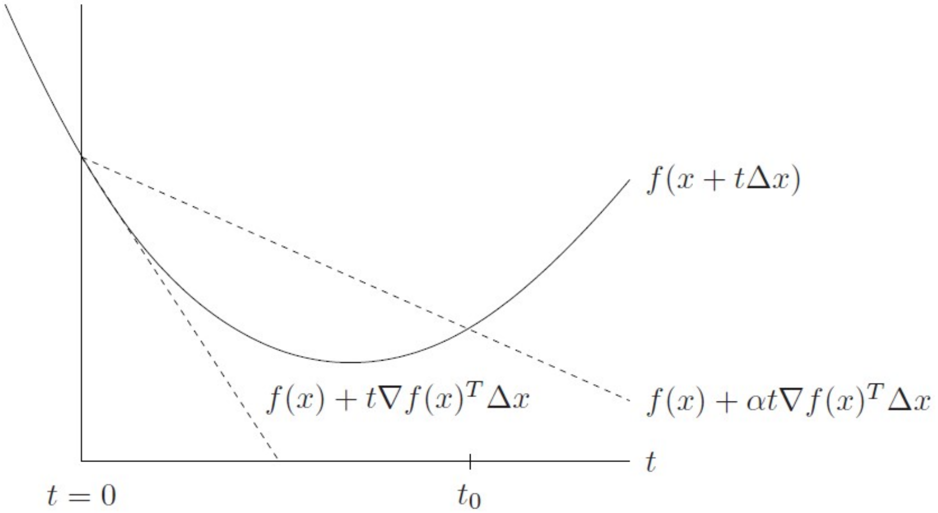

$f(x + t\Delta x)$
$f(x) + t\nabla f(x)^T \Delta x$
$f(x) + \alpha t\nabla f(x)^T \Delta x$
t
$t = 0$
t_0

Figure 1: Backtracking line search (Armijo condition).

The curve shows f (with respect to t), restricted to the line over which we search. The lower dashed line shows the linear approximation of f, and the upper dashed line has a slope a factor of α smaller. The Armijo condition is that f lies below the upper dashed line, i.e., $0 \leq t \leq t_0$.

Under the framework of backtracking line search, there are several conditions that can help us to choose the step length, including Armijo condition, Armijo Goldstein condition and strong Wolfe conditions. More stringent conditions will bring higher computational complexity to the line search method.

2.1.2 Steepest Descent Method

The first-order Taylor approximation of $f(x+v)$ around x is

$$f(x+v) \approx \widehat{f}(x+v) = f(x) + \nabla f(x)^T v.$$

The second term on the righthand side, $\nabla f(x)^T v$, is the directional derivative of f at x in the direction v. Let $|\cdot|$ be any norm on $\mathbf{R}^n$. We define a normalized steepest descent direction (with respect to the norm $|\cdot|$) as

$$\Delta x_{nsd} = \operatorname{argmin}\left\{\nabla f(x)^T v \mid |v| = 1\right\}.$$

If we take the norm $|\cdot|$ to be the Euclidean norm we find that the steepest descent direction is simply the negative gradient, i.e., $\Delta x_{sd} = -\nabla f(x)$. The steepest descent method for the Euclidean norm coincides with the gradient descent method.

2.1.3 Newton's Method

The second-order Taylor approximation (or model) $\widehat{f}$ of f at x is

$$\widehat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v.$$

which is a convex quadratic function of v, and is minimized when $v = \Delta x_{nt}$, where $\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x)$. Thus, the Newton step Δx_{nt} is what should be added to the point x to minimize the second-order approximation of f at x.

2.1.4 Newton's Method with Equality Constraints

Consider a convex optimization problem with equality constraints,

$$\begin{array}{ll} \text{minimize} & f(x), \\ \text{subject to} & Ax = b, \end{array} \tag{3}$$

where $f : \mathbf{R}^n \to \mathbf{R}$ is convex and twice continuously differentiable. Solving the equality constrained optimization problem is therefore equivalent to finding a solution of the KKT equations (4).

$$Ax^\star = b, \quad \nabla f\left(x^\star\right) + A^T \nu^\star = 0. \tag{4}$$

At the feasible point x, we replace the objective with its second-order Taylor approximation near x, to form the problem

$$\begin{array}{rl} \text{minimize} & \widehat{f}(x+v) = f(x) + \nabla f(x)^T v + (1/2) v^T \nabla^2 f(x) v, \\ \text{subject to} & A(x+v) = b. \end{array}$$

with variable v.

From equation (4), the Newton step Δx_{nt} is characterized by

$$\left[\begin{array}{cc} \nabla^2 f(x) & A^T \\ A & 0 \end{array} \right] \left[\begin{array}{c} \Delta x_{nt} \\ w \end{array} \right] = \left[\begin{array}{c} -\nabla f(x) \\ 0 \end{array} \right],$$

where w is the associated optimal dual variable for the quadratic problem. The Newton step is defined only at points for which the KKT matrix is nonsingular.

2.1.5 Interior-point Methods

Consider a convex optimization problem with inequality constraints,

$$\begin{array}{rl} \text{minimize} & f_0(x), \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & Ax = b. \end{array} \tag{5}$$

Interior-point methods solve the problem (5) by applying Newton's method to a sequence of equality constrained problems, or to a sequence of modified versions of the KKT conditions. In *Convex Optimization*, the barrier method was introduced as a particular interior-point algorithm. The barrier method approximately formulate the inequality constrained problem (5) as an equality constrained problem by making the inequality constraints implicit in the objective to which Newton's method can be applied.

2.2 Nondifferentiable Cases

2.2.1 Subgradient

A subgradient of a convex function f at x is any $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T (y - x),$$

for all y.

Optimality Condition

$$f\left(x^{\star}\right) = \min_{x \in \mathbb{R}^n} f(x) \quad \Leftrightarrow \quad 0 \in \partial f\left(x^{\star}\right). \tag{6}$$

Subgradient Method:

Given convex $f : \mathbb{R}^n \rightarrow \mathbb{R}$, subgradient method: initialize $x^{(0)}$, then repeat

$$x^{(k)} = x^{(k-1)} - t_k \cdot g^{(k-1)}, \quad k = 1, 2, 3, \ldots$$

where $g^{(k-1)}$ is any subgradient of f at $x^{(k-1)}$.

Step Size choice

Diminishing step size: choose t_k to satisfy

$$\sum_{k=1}^{\infty} t_k^2 < \infty, \quad \sum_{k=1}^{\infty} t_k = \infty.$$

All step sizes options are pre-specified, not adaptively computed.

2.2.2 Proximal Gradient Descent

Decomposable functions

Suppose

$$f(x) = g(x) + h(x).$$

- g is convex, differentiable, $\operatorname{dom}(g) = \mathbb{R}^n$ - h is convex, not necessarily differentiable

The idea of proximal gradient descent is to make quadratic approximation to g, and calculate h alone.

$$\begin{aligned} x^+ &= \underset{z}{\operatorname{argmin}} \widetilde{g}_t(z) + h(z) \\ &= \underset{z}{\operatorname{argmin}} g(x) + \nabla g(x)^T(z-x) + \frac{1}{2t}\|z-x\|^2 + h(z) \\ &= \underset{z}{\operatorname{argmin}} \frac{1}{2t}\|z-(x-t\nabla g(x))\|^2 + h(z). \end{aligned} \tag{7}$$

Define proximal mapping:

$$\operatorname{prox}_t(x) = \underset{z}{\operatorname{argmin}} \frac{1}{2t}\|x-z\|_2^2 + h(z).$$

Proximal gradient descent: choose initialize $x^{(0)}$, repeat

$$x^{(k)} = \operatorname{prox}_{t_k}\left(x^{(k-1)} - t_k \nabla g\left(x^{(k-1)}\right)\right), \quad k = 1, 2, 3, \ldots$$

2.3 Convergence Performance

2.4 Gradient Descent

For gradient descent method with fixed step t, assume $\nabla f(x)$ is $L-$ smooth, then we have

$$\min_{i=0,\ldots,k} \left\|\nabla f\left(x^{(i)}\right)\right\| \leq \sqrt{\frac{2\left(f\left(x^{(0)}\right) - f\left(x^*\right)\right)}{t(k+1)}},$$

where $f\left(x^*\right)$ represents the minimum value. This result reveals the convergence rate is $\mathcal{O}(\frac{1}{\sqrt{k}})$. If f is convex, then we have

$$f\left(x^{(k)}\right)-f\left(x^{*}\right) \leq \frac{\|x^{(0)}-x^{*}\|^{2}}{2tk}.$$

This result reveals the convergence rate is $\mathcal{O}(\frac{1}{k})$.

2.5 Subgradient Method

Assume that f convex, $\operatorname{dom}(f)=\mathbb{R}^{n}$, and also that f is Lipschitz continuous with constant $G>0$, subgradient method satisfies

$$f\left(x_{best}^{(k)}\right)-f\left(x^{*}\right) \leq \frac{\|x^{(0)}-x^{*}\|^{2}}{2kt}+\frac{G^{2}t}{2}. \tag{8}$$

For a fixed step size t, subgradient will not converge. For diminishing step size, subgradient method satisfies

$$f\left(x_{best}^{(k)}\right)-f\left(x^{*}\right) \leq \frac{\|x^{(0)}-x^{*}\|^{2}+G^{2}\sum_{i=1}^{k}t_{i}^{2}}{2\sum_{i=1}^{k}t_{i}}. \tag{9}$$

This result reveals the convergence rate is $\mathcal{O}(\frac{1}{\sqrt{k}})$.

2.6 Proximal Gradient Descent

With decomposition $f(x)=g(x)+h(x)$, we assume:

- g is convex, differentiable, $\operatorname{dom}(g)=\mathbb{R}^{n}$, and ∇g is Lipschitz continuous with constant $L>0$
- h is convex. $\operatorname{prox}_{t}(x)=\operatorname{argmin}_{z}\left\{\|x-z\|_{2}^{2}/(2t)+h(z)\right\}$ can be evaluated.

  Proximal gradient descent with fixed step size $t \leq 1/L$ satisfies

$$f\left(x^{(k)}\right)-f^{\star} \leq \frac{\|x^{(0)}-x^{\star}\|_{2}^{2}}{2tk}.$$

This result reveals the proximal gradient descent has convergence rate $O(\frac{1}{k})$.

3 Application

3.1 Norm approximation

Norm approximation problem is an unconstrained problem of the form

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

where $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^{m}$ are problem data, $x \in \mathbf{R}^{n}$ is the variable, and $\|\cdot\|$ is a norm on $\mathbf{R}^{m}$. The vector

$$r = Ax-b$$

is called the residual for the problem.

3.1.1 Least-squares approximation

By squaring the objective, we obtain an equivalent problem which is called the least-squares approximation problem,

$$\text{minimize } \|Ax-b\|_2^2 = r_1^2 + r_2^2 + \cdots + r_m^2,$$

where the objective is the sum of squares of the residuals. This problem can be solved analytically by expressing the objective as the convex quadratic function

$$f(x) = x^T A^T A x - 2b^T A x + b^T b.$$

A point x minimizes f if and only if

$$\nabla f(x) = 2A^T A x - 2A^T b = 0.$$

If we assume the columns of A are independent, the least-squares approximation problem has the unique solution $x = \left(A^T A\right)^{-1} A^T b$.

3.1.2 LASSO

Lasso problem can be parametrized as

$$\min_x \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1.$$

where $\lambda \geq 0$. Consider simplified problem with $A = I$:

$$\min_x \frac{1}{2}\|y - x\|^2 + \lambda\|x\|_1.$$

This problem is a convex optimization problem with a nondifferentiable point $x = 0$. Subgradient method and proximal gradient descent method are ready to solve this problem.

The subgradients of $f(x) = \frac{1}{2}\|y-x\|^2 + \lambda\|x\|_1$ are

$$g = x - y + \lambda s.$$

where $s_i = \operatorname{sign}(x_i)$ if $x_i \neq 0$ and $s_i \in [-1, 1]$ if $x_i = 0$.

Then the solution of the problem is $x^\star = S_\lambda(y)$, where S_λ is the soft-thresholding operator:

$$[S_\lambda(y)]_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \leq y_i \leq \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}$$

In proximal gradient descent method, a prox mapping is now

$$\begin{aligned} \operatorname{prox}_t(x) &= \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2t}\|x - z\|_2^2 + \lambda\|z\|_1 \\ &= S_\lambda(x). \end{aligned}$$

Recall $\nabla g(x) = -(y - x)$, hence proximal gradient update is:

$$x^+ = S_\lambda\left(x + t(y - x)\right).$$

3.2 Maximum Likelihood Estimation

Consider a linear measurement model,

$$y_i = a_i^T x + v_i, \quad i = 1, \dots, m$$

where $x \in \mathbf{R}^n$ is a vector of parameters to be estimated, $y_i \in \mathbf{R}$ are the measured or observed quantities, and v_i are the measurement errors or noise. We assume that v_i are independent, identically distributed (IID), with density p on $\mathbf{R}$. The likelihood function is then

$$p_x(y) = \prod_{i=1}^{m} p\left(y_i - a_i^T x\right).$$

So the log-likelihood function is

$$l(x) = \log p_x(y) = \sum_{i=1}^{m} \log p\left(y_i - a_i^T x\right).$$

The ML estimate is any optimal point for the problem

$$\text{maximize} \sum_{i=1}^{m} \log p\left(y_i - a_i^T x\right),$$

with variable x. If the density p is log-concave, this problem is convex.

3.3 Linear discrimination (Support vector classifier)

In linear discrimination, we seek an affine function $f(x) = a^T x - b$ that classifies the points, i.e.,

$$a^T x_i - b > 0, \quad i = 1, \dots, N, \quad a^T y_i - b < 0, \quad i = 1, \dots, M. \tag{10}$$

Equation (10) are feasible if and only if the set of nonstrict linear inequalities

$$a^T x_i - b \geq 1, \quad i = 1, \dots, N, \quad a^T y_i - b \leq -1, \quad i = 1, \dots, M.$$

(in the variables a, b) is feasible.

By relaxing the constraints with nonnegative variables $u_1, \dots, u_N$ and $v_1, \dots, u_M$, and forming the inequalities

$$a^T x_i - b \geq 1, \quad i = 1, \dots, N, \quad a^T y_i - b \leq -1, \quad i = 1, \dots, M. \tag{11}$$

We can think of u_i as a measure of how much the constraint $a^T x_i - b \geq 1$ is violated, and similarly for v_i. Our goal is to find a, b, and sparse nonnegative u and v that satisfy the inequalities (11). As a heuristic for this, we can minimize the sum of the variables u_i and v_i, by solving

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T u + \mathbf{1}^T v, \\ \text{subject to} & a^T x_i - b \geq 1 - u_i, \quad i = 1, \dots, N \\ & a^T y_i - b \leq -\left(1 - v_i\right), \quad i = 1, \dots, M. \end{array} \tag{12}$$

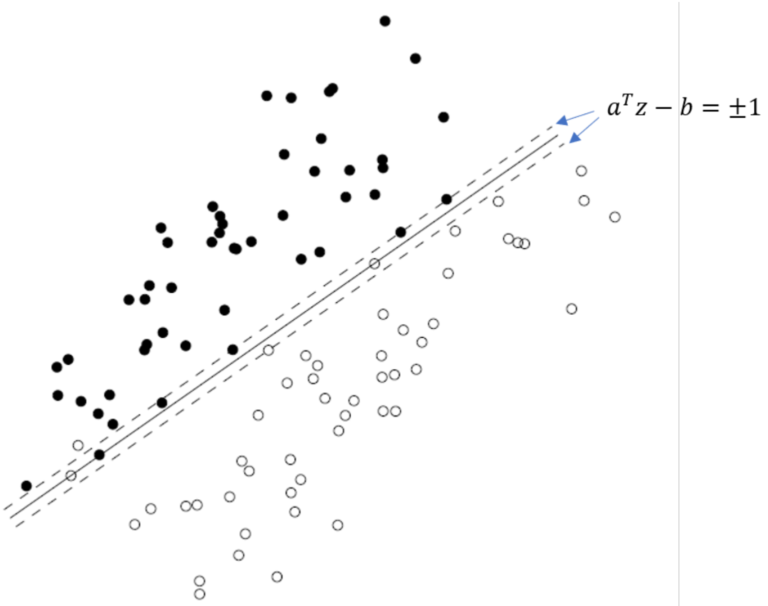


Figure 2: Approximate linear discrimination via linear programming (12).The classifier shown as a solid line.

The objective function in the (12) can be interpreted as a relaxation of the number of points x_i that violate $a^T x_i - b \geq 1$ plus the number of points y_i that violate $a^T y_i - b \leq -1$. In other words, it is a relaxation of the number of points misclassified by the function $a^T z - b$, plus the number of points that are correctly classified but lie in the slab defined by $-1 < a^T z - b < 1$.

More generally, we can consider the trade-off between the number of misclassified points, and the width of the slab $\left\{ z \mid -1 \leq a^T z - b \leq 1 \right\}$, which is given by $\frac{2}{\|a\|_2}$. The standard support vector classifier for the sets $\{x_1, \ldots, x_N\}$,$\{y_1, \ldots, y_M\}$ is defined as the solution of

$$\begin{array}{ll} \text{minimize} & \|a\|_2 + \gamma \left(1^T u + 1^T v \right), \\ \text{subject to} & a^T x_i - b \geq 1 - u_i, \quad i = 1, \ldots, N \\ & a^T y_i - b \leq -\left(1 - v_i\right), \quad i = 1, \ldots, M \\ & u \succeq 0, \quad v \succeq 0. \end{array} \tag{13}$$

The first term is proportional to the inverse of the width of the slab defined by $-1 \leq a^T z - b \leq 1$. The second term has the same interpretation as above. The parameter γ, which is positive, gives the relative weight of the number of misclassified points (which we want to minimize), compared to the width of the slab (which we want to maximize).