17 Constrained Optimization

best Chapter 15
best frequently has constraints
nutrition
groceries $\vec{g} \geq 0$
prices $\vec{p}$
price $\min_{\vec{g}} \vec{g} \cdot \vec{p}$
minimum health requirements $\vec{m}$
nutrition value **N**
$\mathbf{N} \cdot \vec{g} \geq \vec{m}$
defines linear program, LP
price may be a function of quantity, not linear
quadratic objective, quadratic program, QP
general case mathematical program
portfolios, routing airplanes, running a factory
program as plan, not computer program, but can be same
electrical networks [Dennis, 1958]
routing [Kelly, 1991, Papadimitriou & Steiglitz, 1998]
flow control [Low *et al.*, 2002]
layering [Chiang *et al.*, 2007]
sorting
variables $\vec{x}$, objective minimize $f(\vec{x})$, constraints $\vec{c}(\vec{x})$
max = -min
slack variables to convert inequality to equality

$$c(\vec{x}) \geq 0 \tag{17.1}$$

replace with

$$\begin{aligned} c(\vec{x}) - s &= 0 \\ s &\geq 0 \end{aligned} \tag{17.2}$$

combinatorial x equals 1 or -1 can be relaxed as algebraic constraint $(x^2 - 1)^2 = 0$
L1 norm

$$|\vec{x}|_1 = \sum_i |x_i| \tag{17.3}$$

compressed sensing, sparsity
non-differentiable
[Schmidt *et al*., 2007]
can be relaxed

$$\begin{aligned} |x| &\approx |x|_\alpha \\ &= \frac{1}{\alpha}\left[\log\left(1+e^{-\alpha x}\right) + \log\left(1+e^{\alpha x}\right)\right] \end{aligned} \tag{17.4}$$

$$\frac{d|x|_\alpha}{dx} = \frac{1}{1+e^{-\alpha x}} - \frac{1}{1+e^{\alpha x}} \tag{17.5}$$

$$\frac{d^2|x|_\alpha}{dx^2} = \frac{2\alpha e^{\alpha x}}{(1+e^{\alpha x})^2} \tag{17.6}$$

minimize for increasing α

17.1 LAGRANGE MULTIPLIERS

single equality constraint $c(\vec{x}) = 0$
step in direction $\vec{\delta}$ to minimize f while satisfying the constraint

$$\begin{aligned} 0 &= c(\vec{x}+\vec{\delta}) \\ &\approx c(\vec{x}) + \nabla c \cdot \vec{\delta} \\ &= \nabla c \cdot \vec{\delta} \end{aligned} \tag{17.7}$$

step also minimizes f

$$\begin{aligned} 0 &> f(\vec{x}+\vec{\delta}) - f(\vec{x}) \\ &\approx f(\vec{x}) + \nabla f \cdot \vec{\delta} - f(\vec{x}) \\ &= \nabla f \cdot \vec{\delta} \end{aligned} \tag{17.8}$$

if $\nabla c(\vec{x})$ and $\nabla f(\vec{x})$ aligned not possible to find a direction, hence $\vec{x}$ is a local minimizer
define *Lagrangian*

$$\mathcal{L} = f(\vec{x}) - \lambda c(\vec{x}) \tag{17.9}$$

solve for

$$\begin{aligned} 0 &= \nabla \mathcal{L} \\ &= \nabla f - \lambda \nabla c \end{aligned} \tag{17.10}$$

multiple constraints
linear combination

$$\nabla f(\vec{x}) = \sum_i \lambda_i \nabla c_i(\vec{x}) \tag{17.11}$$

$$f(\vec{x}) = \sum_i \lambda_i c_i(\vec{x}) \tag{17.12}$$

solving gives $\vec{x}(\vec{\lambda})$, substitute into constraints to find $\vec{\lambda}$
inequality constraint

$$\begin{aligned} 0 &\leq c(\vec{x} + \vec{\delta}) \\ &\approx c(\vec{x}) + \nabla c \cdot \vec{\delta} \end{aligned} \tag{17.13}$$

if constraint not active ($c > 0$), can just do gradient descent $\vec{\delta} = -\alpha \nabla f$
for an active constraint $\nabla f \cdot \vec{\delta} < 0$ and $\nabla c \cdot \vec{\delta} \geq 0$
define half-planes
no intersection if point in same direction $\nabla f = \lambda \nabla c$
same condition, but now $\lambda \geq 0$

17.2 OPTIMALITY

first-order condition
equality constraints $c_i(\vec{x}), i \in \mathcal{E}$
inequality constraints $c_i(\vec{x}), i \in \mathcal{I}$
inactive constraint $\lambda_i = 0$
complementarity: $\lambda_i c_i = 0$: Lagrange multiplier only non-zero when constraint is active, otherwise reduces to gradient descent

$$\begin{aligned} \nabla_{\vec{x}} \mathcal{L}(\vec{x}, \vec{\lambda}) &= 0 \\ c_i(\vec{x}) &= 0 \quad (i \in \mathcal{E}) \\ c_i(\vec{x}) &\geq 0 \quad (i \in \mathcal{I}) \\ \lambda_i &\geq 0 \quad (i \in \mathcal{I}) \\ \lambda_i c_i(x) &= 0 \end{aligned} \tag{17.14}$$

Karush-Kuhn-Tucker (KKT) conditions
necessary, not sufficient
second order condition: positive definite Lagrangian Hessian
sensitivity

replace $c(x) = 0$ with $c(x) = \epsilon$
minimizer $\vec{x}$ goes to $\vec{x}_\epsilon$

$$\begin{aligned} f(\vec{x}_\epsilon) - f(\vec{x}) &\approx \nabla f \cdot (\vec{x}_\epsilon - \vec{x}) \\ &= \lambda \nabla c \cdot (\vec{x}_\epsilon - \vec{x}) \\ &\approx \lambda \left(c(\vec{x}_\epsilon) - c(\vec{x})\right) \\ &= \lambda \epsilon \\ \frac{df}{d\epsilon} &= \lambda \end{aligned} \tag{17.15}$$

shadow prices: change in utility per change in constraint
multi-objective
boundary where not possible to improve one constraint without making others worse
defines Pareto frontier
can combine in multi-objective function with relative weights

17.3 SOLVERS

17.3.1 Penalty

$$\mathcal{F} = f(\vec{x}) + \frac{\mu}{2} \sum_i c_i^2(\vec{x}) \tag{17.16}$$

$$\frac{\partial \mathcal{F}}{\partial x_j} = \frac{\partial f}{\partial x_j} + \mu \sum_i c_i \frac{\partial c_i}{\partial x_j} \tag{17.17}$$

$$\mathcal{L} = f(\vec{x}) - \sum_i \lambda_i c_i(\vec{x}) \tag{17.18}$$

$$\frac{\partial \mathcal{L}}{\partial x_j} = \frac{\partial f}{\partial x_j} - \sum_i \lambda_i \frac{\partial c_i}{\partial x_j} \tag{17.19}$$

effectively taking $c_i = -\lambda_i / \mu$
solving a different problem
constraint driven to 0 as $\mu \to \infty$
small μ may be unbounded
large μ may be ill-conditioned
nonsmooth penalty

$$\mathcal{F} = f(\vec{x}) + \mu \sum_{i \in E} |c_i(\vec{x})| + \mu \sum_{i \in I} [c_i(\vec{x})]_- \tag{17.20}$$

can be exact for large μ [Nocedal & Wright, 2006]
non-differentiable
approximate (17.4)

17.3.2 Augmented Lagrangian

augmented Lagrangian

$$\mathcal{L} = f(\vec{x}) - \sum_i \lambda_i c_i(\vec{x}) + \frac{\mu}{2} \sum_i c_i^2(\vec{x}) \tag{17.21}$$

$$\frac{\partial \mathcal{L}}{\partial x_j} = \frac{\partial f}{\partial x_j} - \sum_i \lambda_i \frac{\partial c_i}{\partial x_j} + \mu \sum_i c_i \frac{\partial c_i}{\partial x_j} \tag{17.22}$$

$\lambda_i^* = \lambda_i - \mu c_i$
$c_i = (\lambda_i - \lambda_i^*)/\mu$
vanishes much faster, as Lagrange multiplier estimates converge
$\lambda_i^{(n+1)} = \lambda_i^{(n)} - \mu c_i$
minimize $\vec{x}$, update λ, increase μ

17.3.3 Interior Point

interior point
basis largest, most efficient solvers
directly solve KKT system of equations

$$\begin{aligned} \min_{\vec{x}} f(\vec{x}) & \\ c_i(\vec{x}) = 0 \quad & (i \in \mathcal{E}) \\ c_i(\vec{x}) - s_i = 0 \quad & (i \in \mathcal{I}) \\ s_i \geq 0 & \end{aligned} \tag{17.23}$$

perturb from boundary

$$\begin{aligned} \nabla f - \sum_i \lambda_i \nabla c_i(\vec{x}) = 0 & \\ c_i(\vec{x}) = 0 \quad & (i \in \mathcal{E}) \\ c_i(\vec{x}) - s_i = 0 \quad & (i \in \mathcal{I}) \\ \lambda_i s_i = \mu \quad & (i \in \mathcal{I}) \end{aligned} \tag{17.24}$$

iterate Newton step on system, decrease μ
same as barrier method

$$\begin{aligned} \min_{\vec{x},\vec{s}} f(x) - \mu \sum_i \log s_i \quad & (i \in \mathcal{I}) \\ c_i(\vec{x}) = 0 \quad & (i \in \mathcal{E}) \\ c_i(\vec{x}) - s_i = 0 \quad & (i \in \mathcal{I}) \end{aligned} \tag{17.25}$$

KKT condition for s_i

$$\mu \frac{1}{s_i} - \lambda_i = 0 \tag{17.26}$$

$$\lambda_i s_i = \mu \tag{17.27}$$

17.4 CONVEXITY

Convexity watershed [Rockafellar, 1993]

convex semidefinite program [Vandenberghe & Boyd, 1996, Boyd & Vandenberghe, 2004]

definitions

solvers

"non-iterative"

17.4.1 Kernels

kernel k

terminology from functional analysis of integral equations

here symmetric positive definite functions

K kernel matrix $K_{ij} = k(\vec{x}_i, \vec{x}_j)$

e.g. Gaussian kernel

$$K_{ij} = e^{-|\vec{x}_i - \vec{x}_j|^2/(2\sigma^2)} \tag{17.28}$$

K symmetric positive definite $\Rightarrow \mathbf{K} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$

V columns eigenvectors

ith row $\vec{v}_i$

$\mathbf{\Lambda}$ diagonal eigenvalues

$$\begin{aligned} k(\vec{x}_i, \vec{x}_j) &= K_{ij} \\ &= \left(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\right)_{ij} \\ &= \vec{v}_i \cdot \mathbf{\Lambda}\vec{v}_j \\ &= \left(\sqrt{\mathbf{\Lambda}}\ \vec{v}_i\right) \cdot \left(\sqrt{\mathbf{\Lambda}}\ \vec{v}_j\right) \\ &\equiv \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_j) \end{aligned} \tag{17.29}$$

symmetric positive definite kernel implies dot product

$$\begin{aligned} \sum_{ij} \alpha_i k(\vec{x}_i, \vec{x}_j) \alpha_j &= \sum_{ij} \alpha_i \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_j) \alpha_j \\ &= \left(\sum_i \alpha_i \vec{\varphi}(\vec{x}_i)\right) \cdot \left(\sum_j \alpha_j \vec{\varphi}(\vec{x}_j)\right) \end{aligned}$$

$$\begin{aligned} &= \left| \sum_i \alpha_i \vec{\varphi}(\vec{x}_i) \right|^2 \\ &\geq 0 \end{aligned} \tag{17.30}$$

dot product implies symmetric positive definite
Mercer's thm for functions [Schölkopf & Smola, 2002]
generalize dot product, possibly infinite dimensional
map feature vector to symmetric positive definite kernel $\Phi(\vec{x}) = k(\cdot, \vec{x})$
$[\Phi(\vec{x})](\vec{x}') = k(\vec{x}', \vec{x})$
define vector space

$$f(\cdot) = \sum_i \alpha_i k(\cdot, x_i) \tag{17.31}$$

define inner product with $g(\cdot) = \sum_j \beta_j k(\cdot, x_j)$

$$\begin{aligned} \langle f(\cdot), g(\cdot) \rangle &= \left\langle \sum_i \alpha_i k(\cdot, x_i), \sum_j \beta_j k(\cdot, x'_k) \right\rangle \\ &= \sum_{ij} \alpha_i \left\langle k(\cdot, x_i), k(\cdot, x'_j) \right\rangle \beta_j \\ &= \sum_{ij} \alpha_i k(x_i, x'_j) \beta_j \end{aligned} \tag{17.32}$$

$$\begin{aligned} \langle \Phi(\vec{x}), \Phi(\vec{x}') \rangle &= \langle k(\cdot, \vec{x}), k(\cdot, \vec{x}') \rangle \\ &= k(\vec{x}, \vec{x}') \end{aligned} \tag{17.33}$$

giving dot product representation

$$\begin{aligned} \langle k(\cdot, \vec{x}), f(\cdot) \rangle &= \left\langle k(\cdot, \vec{x}), \sum_i \alpha_i k(\cdot, \vec{x}_i) \right\rangle \\ &= \sum_i \alpha_i \langle k(\cdot, \vec{x}), k(\cdot, \vec{x}_i) \rangle \\ &= \sum_i \alpha_i k(\vec{x}, \vec{x}_i) \\ &= f(\vec{x}) \end{aligned} \tag{17.34}$$

reproducing property

$$|f(\vec{x})|^2 = |\langle k(\cdot, \vec{x}), f \rangle|^2 \leq k(\vec{x}, \vec{x}) \langle f(\cdot), f(\cdot) \rangle \tag{17.35}$$

f vanishes if dot product vanishes

$$\langle f(\cdot), f(\cdot) \rangle = \sum_{ij} \alpha_i k(x_i, x_j) \alpha_j \geq 0 \tag{17.36}$$

dot product non-negative
with norm $\sqrt{\langle f(\cdot), f(\cdot)\rangle}$ defines Reproducing Kernel Hilbert Space (RKHS) [Berlinet & Thomas-Agnan, 2004]
representer theorem
SDP hierarchy of convex relaxations [Parrilo, 2003]
SDP for combinatorial optimization [Goemans & Williamson, 1995, Goemans & Williamson, 2004]
SDP for global polynomial minimization [Lasserre, 2002, Lasserre, 2006]

17.4.2 Support Vector Machines

[Vapnik, 1998, Burges, 1998]

17.4.2.1 Classification

classification
$\{\vec{x}_i, y_i\}$
linear classifier

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b \tag{17.37}$$

$f(\vec{x}) = 0$ defines hyperplane
$f \geq 1$ for $y = 1$, $f \leq -1$ for $y = -1$
combine

$$y_i \left(\vec{w} \cdot \vec{x}_i + b\right) \geq 1 \tag{17.38}$$

margin $|\vec{x}| = |\vec{w}|^{-1}$
maximize margin

$$\begin{aligned} &\min_{\vec{w}} \frac{1}{2} \vec{w} \cdot \vec{w} \\ \text{subject to } \quad &y_i \left(\vec{w} \cdot \vec{x}_i + b\right) \geq 1 \end{aligned} \tag{17.39}$$

(1/2 for algebraic convenience)
to find $\vec{w}$, search in size of feature vector D
Lagrangian

$$\mathcal{L} = \frac{1}{2} \vec{w} \cdot \vec{w} - \sum_i \lambda_i \left[y_i \left(\vec{w} \cdot \vec{x}_i + b\right) - 1\right] \tag{17.40}$$

set

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial \vec{w}} \\ &= \vec{w} - \sum_i \lambda_i y_i \vec{x}_i \\ \Rightarrow \vec{w} &= \sum_i \lambda_i y_i \vec{x}_i \end{aligned} \tag{17.41}$$

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial b} \\ &= \sum_i \lambda_i y_i \end{aligned} \tag{17.42}$$

substitute

$$\begin{aligned} \mathcal{L} &= \frac{1}{2} \sum_{i,j} \lambda_i y_i \vec{x}_i \cdot \vec{x}_j y_j \lambda_j - \sum_{i,j} \lambda_i y_i \vec{x}_i \cdot \vec{x}_j y_j \lambda_j - b \sum_i \lambda_i y_i + \sum_i \lambda_i \\ &= \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i y_i \vec{x}_i \cdot \vec{x}_j y_j \lambda_j \\ &= \sum_i \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i y_i G_{ij} y_j \lambda_j \end{aligned} \tag{17.43}$$

Gram matrix $G_{ij} = \vec{x}_i \cdot \vec{x}_j$
dual

$$\begin{aligned} \min_{\vec{\lambda}} \frac{1}{2} \sum_{i,j} \lambda_i y_i G_{ij} y_j \lambda_j - \sum_i \lambda_i \\ \text{subject to} \qquad \lambda_i > 0 \\ \sum_i \lambda_i y_i = 0 \end{aligned} \tag{17.44}$$

to find $\vec{\lambda}$, search in number of points N
advantageous for huge feature vectors
QP
dual solution $\vec{\lambda} \rightarrow$ primal solution $\vec{w} \rightarrow$ KKT b

$$\lambda_i \left[y_i \left(\vec{w} \cdot \vec{x} + b \right) - 1 \right] = 0 \tag{17.45}$$

$$\begin{aligned} f(\vec{x}) &= \vec{w} \cdot \vec{x} + b \\ &= \sum_i \lambda_i y_i \vec{x}_i \cdot \vec{x} + b \end{aligned} \tag{17.46}$$

λ_i nonzero only for constraint equality: define support vectors at boundary
sparse sum over support vectors
nonlinear

$$f(\vec{x}) = \vec{W} \cdot \vec{\varphi}(\vec{x}) + b \tag{17.47}$$

$$\mathcal{L} = \sum_i \lambda_i - \sum_{i,j} \lambda_i y_i \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_i) y_j \lambda_j \tag{17.48}$$

nonlinear classifier with kernel trick

$$\mathcal{L} = \sum_i \lambda_i - \sum_{i,j} \lambda_i y_i K_{ij} y_j \lambda_j \tag{17.49}$$

soft margin

$$y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) \geq 1 - s_i \tag{17.50}$$

slack s_i allows violation

$$\begin{aligned} \min \frac{1}{2} \vec{W} \cdot \vec{W} + \lambda_s \sum_i s_i \\ \text{subject to } \; y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) \geq 1 - s_i \\ s_i \geq 0 \end{aligned} \tag{17.51}$$

Lagrangian

$$\mathcal{L} = \frac{1}{2} \vec{W} \cdot \vec{W} + \lambda_s \sum_i s_i - \sum_i \lambda_i \left[y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) - 1 + s_i \right] - \sum_i \lambda_{s,i} s_i \tag{17.52}$$

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial s_i} \\ &= \lambda_s - \lambda_i - \lambda_{s,i} \\ \Rightarrow \lambda_s &= \lambda_i + \lambda_{s,i} \end{aligned} \tag{17.53}$$

$\lambda_{s,i} \geq 0$, $\lambda_i \geq 0$ therefore $0 \leq \lambda_i \leq \lambda_s$
same problem, now with upper bound on λ_i
KKT

$$\lambda_i \left[y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) - 1 + s_i \right] = 0 \tag{17.54}$$

$$\lambda_{s,i} s_i = 0 \tag{17.55}$$

$$(\lambda_s - \lambda_i) \, s_i = 0 \tag{17.56}$$

three cases
$\lambda_i = 0$ OK
$\lambda_i = \lambda_s$ implies $s_i \neq 0$ wrong side
$0 < \lambda_i < \lambda_s$ implies $s_i = 0$ support vector
M-M

$$\begin{aligned} \min \frac{1}{2}\vec{W}\cdot\vec{W} + b^2 + \lambda_s \sum_i s_i \\ \text{subject to } \; y_i\left(\vec{W}\cdot\vec{\varphi}(\vec{x}_i)+b\right) \geq 1 - s_i \\ s_i \geq 0 \end{aligned} \tag{17.57}$$

include b in norm
[Mangasarian & Musicant, 1999]
Lagrangian

$$\mathcal{L} = \frac{1}{2}\vec{W}\cdot\vec{W} + b^2 + \lambda_s \sum_i s_i - \sum_i \lambda_i \left[y_i\left(\vec{W}\cdot\vec{\varphi}(\vec{x}_i)+b\right) - 1 + s_i\right] - \sum_i \lambda_{s,i} s_i \tag{17.58}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \;\Rightarrow\; b = \frac{1}{2}\sum_i \lambda_i y_i \tag{17.59}$$

$$\mathcal{L} = \sum_i \lambda_i - \sum_{i,j} \lambda_i y_i (K_{ij} + 1) y_j \lambda_j \tag{17.60}$$

no equality constraint
least-squares
[Suykens *et al*., 2002]

$$\begin{aligned} \min \frac{1}{2}\vec{W}\cdot\vec{W} + \lambda_s \frac{1}{2}\sum_i s_i^2 \\ \text{subject to } \; y_i\left(\vec{W}\cdot\vec{\varphi}(\vec{x}_i)+b\right) = 1 - s_i \end{aligned} \tag{17.61}$$

error target rather than threshold
equality constraint, sign doesn't matter

$$\frac{\partial \mathcal{L}}{\partial s_i} = 0 \Rightarrow \lambda_i = \lambda_s s_i \tag{17.62}$$

linear KKT

$$\begin{bmatrix} 0 & y_1 & \cdots & y_N \\ y_1 & & \vdots & \\ \vdots & \cdots & y_i K_{ij} y_j + 1/\lambda_s & \cdots \\ y_N & & \vdots & \end{bmatrix} \begin{bmatrix} b \\ \lambda_1 \\ \vdots \\ \lambda_N \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{17.63}$$

lose sparsity
multiple classes [Vapnik, 1998]
class k

$$f_k(\vec{x}) = \vec{W}_k \cdot \vec{\varphi}(\vec{x}) + b_k \tag{17.64}$$

$$\text{argmax}\{\vec{W}_1 \cdot \vec{\varphi}(\vec{x}) + b_1, \ldots, \vec{W}_k \cdot \vec{\varphi}(\vec{x}) + b_k, \ldots\} \tag{17.65}$$

could construct each separately
add constraint and find jointly

$$\vec{W}_k \cdot \vec{\varphi}(\vec{x}_k) + b_k - \vec{W}_m \cdot \vec{\varphi}(\vec{x}_k) + b_m \geq 1 - s_{k,i} \tag{17.66}$$

$m \neq k$

17.4.2.2 Regression

regression [Smola & Schölkopf, 2004]

$$f(\vec{x}) = \vec{W} \cdot \vec{\varphi}(\vec{x}) + b \tag{17.67}$$

sparsity $\Rightarrow$ ignore errors $< \epsilon$
like SVD, minimize magnitude $\vec{W}$
primal

$$\begin{aligned} \min \frac{1}{2}\vec{W} \cdot \vec{W} + \lambda_s \sum_i (s_i^+ + s_i^-) \\ \text{subject to } \; y_i - \vec{W} \cdot \vec{\varphi}(\vec{x}_i) - b \geq \epsilon + s_i^+ \\ \vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b - y_i \geq \epsilon + s_i^- \\ s_i^+, s_i^- \geq 0 \end{aligned} \tag{17.68}$$

Lagrangian

$$\begin{aligned} \mathcal{L} = \frac{1}{2}\vec{W} \cdot \vec{W} + \lambda_s \sum_i \left(s_i^+ + s_i^-\right) - \sum_i \left(\lambda_{s,i}^+ s_i^+ + \lambda_{s,i}^- s_i^-\right) \\ - \sum_i \lambda_i^+ \left(\epsilon + s_i^+ - y_i + \vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b\right) \\ - \sum_i \lambda_i^- \left(\epsilon + s_i^- + y_i - \vec{W} \cdot \vec{\varphi}(\vec{x}_i) - b\right) \end{aligned} \tag{17.69}$$

$$\frac{\partial \mathcal{L}}{\partial \vec{W}} = 0 \;\Rightarrow\; \vec{W} = \sum_i \left(\lambda_i^+ - \lambda_i^-\right) \vec{\varphi}(\vec{x}_i) \tag{17.70}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \;\Rightarrow\; \sum_i \left(\lambda_i^+ - \lambda_i^-\right) = 0 \tag{17.71}$$

$$\frac{\partial \mathcal{L}}{\partial s_i^+} = 0 \;\Rightarrow\; \lambda_s = \lambda_{s,i}^+ + \lambda_i^+ \tag{17.72}$$

$$\frac{\partial \mathcal{L}}{\partial s_i^-} = 0 \;\Rightarrow\; \lambda_s = \lambda_{s,i}^- + \lambda_i^- \tag{17.73}$$

dual

$$\begin{aligned} \max \; & -\frac{1}{2}\sum_{i,j}\left(\lambda_i^+ - \lambda_i^-\right)\left(\lambda_j^+ - \lambda_i^-\right)K_{ij} \\ & -\epsilon\sum_i\left(\lambda_i^+ + \lambda_i^-\right) + \sum_i y_i\left(\lambda_i^+ - \lambda_i^-\right) \\ & \text{subject to} \quad \sum_i\left(\lambda_i^+ - \lambda_i^-\right) = 0 \\ & 0 \le \lambda_i^+, \lambda_i^- \le \lambda_s \end{aligned} \tag{17.74}$$

$$\begin{aligned} f(\vec{x}) &= \vec{W}\cdot\vec{\varphi}(\vec{x}) + b \\ &= \sum_i\left(\lambda_i^+ - \lambda_i^-\right)K(\vec{x},\vec{x}_i) + b \end{aligned} \tag{17.75}$$

17.4.2.3 Clustering

unsupervised learning

clustering [Ben-Hur *et al*., 2001]

$$|\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0|^2 \le R^2 \tag{17.76}$$

$$|\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0|^2 \le R^2 + s_i \tag{17.77}$$

want smallest hypersphere: minimize magnitude R

$$\mathcal{L} = R^2 + \lambda_s\sum_i s_i - \sum_i \lambda_i\left[R^2 + s_i - |\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0|^2\right] - \sum_i \lambda_{s,i}s_i \tag{17.78}$$

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial R} &= 2R - 2R\sum_i\lambda_i = 0 \\ &\Rightarrow \sum_i\lambda_i = 1 \end{aligned} \tag{17.79}$$

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\vec{\varphi}_0} &= \frac{\partial}{\partial\vec{\varphi}_0}\sum_i\lambda_i\left(\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0\right)\cdot\left(\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0\right) \\ &= \sum_i\lambda_i\left(-2\vec{\varphi}(\vec{x}_i) + 2\vec{\varphi}_0\right) \\ &= 0 \\ \Rightarrow \vec{\varphi}_0 &= \sum_i\lambda_i\vec{\varphi}(\vec{x}_i) \end{aligned} \tag{17.80}$$

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial s_i} &= \lambda_s - \lambda_i - \lambda_{s,i} = 0 \\ \Rightarrow \lambda_{s,i} &= \lambda_s - \lambda_i \end{aligned} \tag{17.81}$$

KKT

$$\lambda_{s,i} s_i = 0 \tag{17.82}$$

$$\lambda_i \left[R^2 + s_i - |\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0|^2 \right] = 0 \tag{17.83}$$

support vectors
substitute

$$\mathcal{L} = \sum_i \lambda_i \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_i) - \sum_{i,j} \lambda_i \lambda_j \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_j) \tag{17.84}$$

$$\mathcal{L} = \sum_i \lambda_i K_{ii} - \sum_{i,j} \lambda_i \lambda_j K_{ij} \tag{17.85}$$

$$0 \leq \lambda_i \leq \lambda_s \tag{17.86}$$

radius

$$R(\vec{x}_i) = K_{ii} - 2 \sum_j \lambda_j K_{ij} + \sum_{i,j} \lambda_i \lambda_j K_{ij} \tag{17.87}$$

boundary radius of support vectors
cluster assignment line $> R$ adjacency matrix

17.5 SELECTED REFERENCES

[Nocedal & Wright, 2006] Nocedal, Jorge, & Wright, Stephen J. (2006). *Numerical Optimization*. 2nd edn. New York: Springer.

Unusually clear coverage of a field full of unusually opaque books.

[Boyd & Vandenberghe, 2004] Boyd, Stephen, & Vandenberghe, Lieven. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

Authoritative survey of convex problems and solutions.

17.6 PROBLEMS

(16.1) Given a point (x_0, y_0), analytically find the closest point on the line $y = ax + b$ by minimizing the distance $d^2 = (x_0 - x)^2 + (y_0 - y)^2$ subject to the constraint $y - ax - b = 0$.

(16.2) Consider a set of N nodes that has each measured a quantity x_i. The goal is to find the best estimate $\bar{x}$ by minimizing

$$\min_{\bar{x}} \sum_{i=1}^{N} (\bar{x} - x_i)^2 \quad , \tag{17.88}$$

however each node i can communicate only with nodes j in its neighborhood $j \in \mathcal{N}(i)$. This can be handled by having each node obtain a local estimate $\bar{x}_i$, and introducing a consistency constraint $c_{ij} = \bar{x}_i - \bar{x}_j = 0 \; \forall \; j \in \mathcal{N}(i)$.

(*a*) What is the Lagrangian?

(*b*) Find an update rule for the estimates $\bar{x}_i$ by evaluating where the gradient of the Lagrangian vanishes.

(*c*) Find an update rule for the Lagrange multipliers by taking a Newton root-finding step on their associated constraints.

(16.3) Sorting can be written in terms of a permutation matrix $\mathbf{P}$ as $\vec{s} = \mathbf{P} \cdot \vec{u}$, where $\vec{u}$ is a vector of unsorted numbers, $\vec{s}$ are the sorted numbers, and each row and column of $\mathbf{P}$ has one 1 and the rest of the elements are 0. Defining the vector $\vec{n}$ to be $\{1, 2, \ldots\}$, sorting can be done by maximizing $\vec{n} \cdot \vec{s}$. Solve this as a constrained optimization for a vector of random numbers.