18 Convex Optimization

Convexity watershed [Rockafellar, 1993]
convex semidefinite program [Vandenberghe & Boyd, 1996, Boyd & Vandenberghe, 2004]

18.1 COMPRESSED SENSING

Nyquist
measure, then compress
compress, then measure
sparsity
linear, nonadaptive
random matrix
null space property(NSP), restricted isometry property (RIP)
Gaussian random matrix, variance 1/m
random partial Fourier matrix, m rows
min $|x|$ Ax = y
L0, support, combinatorial, NP-hard
L1, linear program, constrained optimization
hyperplane, norm intersection
relaxation
[Fornasier & Rauhut, 2011]
[Donoho, 2006]
[Candès *et al*., 2006; Candes *et al*., 2006; Candes & Tao, 2006]
[Kim *et al*., 2007]
[Yang *et al*., 2010]

18.2 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

$$K_{ij} = e^{-|\vec{x}_i - \vec{x}_j|^2/(2\sigma^2)} \tag{18.1}$$

$\mathbf{K}$ symmetric positive definite $\Rightarrow \mathbf{K} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$
$\mathbf{V}$ columns eigenvectors
ith row $\vec{v}_i$
$\mathbf{\Lambda}$ diagonal eigenvalues

$$\begin{aligned} k(\vec{x}_i, \vec{x}_j) &= K_{ij} \\ &= \left(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\right)_{ij} \\ &= \vec{v}_i \cdot \mathbf{\Lambda}\vec{v}_j \\ &= \left(\sqrt{\mathbf{\Lambda}}\,\vec{v}_i\right) \cdot \left(\sqrt{\mathbf{\Lambda}}\,\vec{v}_j\right) \\ &\equiv \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_j) \end{aligned} \tag{18.2}$$

symmetric positive definite kernel implies dot product

$$\begin{aligned} \sum_{ij} \alpha_i k(\vec{x}_i, \vec{x}_j)\alpha_j &= \sum_{ij} \alpha_i \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_j)\alpha_j \\ &= \left(\sum_i \alpha_i \vec{\varphi}(\vec{x}_i)\right) \cdot \left(\sum_j \alpha_j \vec{\varphi}(\vec{x}_j)\right) \\ &= \left|\sum_i \alpha_i \vec{\varphi}(\vec{x}_i)\right|^2 \\ &\geq 0 \end{aligned} \tag{18.3}$$

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

$$f(\cdot) = \sum_i \alpha_i k(\cdot, x_i) \tag{18.4}$$

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

$$\begin{aligned} \langle f(\cdot), g(\cdot)\rangle &= \left\langle \sum_i \alpha_i k(\cdot, x_i), \sum_j \beta_j k(\cdot, x'_k) \right\rangle \\ &= \sum_{ij} \alpha_i \left\langle k(\cdot, x_i), k(\cdot, x'_j)\right\rangle \beta_j \\ &= \sum_{ij} \alpha_i k(x_i, x'_j)\beta_j \end{aligned} \tag{18.5}$$

$$\begin{aligned}\langle \Phi(\vec{x}), \Phi(\vec{x}') \rangle &= \langle k(\cdot, \vec{x}), k(\cdot, \vec{x}') \rangle \\ &= k(\vec{x}, \vec{x}') \end{aligned} \tag{18.6}$$

giving dot product representation

$$\begin{aligned}\langle k(\cdot, \vec{x}), f(\cdot) \rangle &= \left\langle k(\cdot, \vec{x}), \sum_i \alpha_i k(\cdot, \vec{x}_i) \right\rangle \\ &= \sum_i \alpha_i \langle k(\cdot, \vec{x}), k(\cdot, \vec{x}_i) \rangle \\ &= \sum_i \alpha_i k(\vec{x}, \vec{x}_i) \\ &= f(\vec{x}) \end{aligned} \tag{18.7}$$

reproducing property

$$|f(\vec{x})|^2 = |\langle k(\cdot, \vec{x}), f \rangle|^2 \leq k(\vec{x}, \vec{x}) \, \langle f(\cdot), f(\cdot) \rangle \tag{18.8}$$

f vanishes if dot product vanishes

$$\langle f(\cdot), f(\cdot) \rangle = \sum_{ij} \alpha_i k(x_i, x_j) \alpha_j \geq 0 \tag{18.9}$$

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

18.3 SUPPORT VECTOR MACHINES

machine learning
[Vapnik, 1998, Burges, 1998]
loss function

18.3.1 Classification

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

$$f(\vec{x}) = \vec{w} \cdot \vec{x} + b \tag{18.10}$$

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

$$y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1 \tag{18.11}$$

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

$$\begin{aligned} &\min_{\vec{w}} \frac{1}{2}\vec{w}\cdot\vec{w} \\ \text{subject to } \quad &y_i(\vec{w}\cdot\vec{x}_i + b) \geq 1 \end{aligned} \tag{18.12}$$

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

$$\mathcal{L} = \frac{1}{2}\vec{w}\cdot\vec{w} - \sum_i \lambda_i [y_i(\vec{w}\cdot\vec{x}_i + b) - 1] \tag{18.13}$$

set

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial \vec{w}} \\ &= \vec{w} - \sum_i \lambda_i y_i \vec{x}_i \\ \Rightarrow \vec{w} &= \sum_i \lambda_i y_i \vec{x}_i \end{aligned} \tag{18.14}$$

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial b} \\ &= \sum_i \lambda_i y_i \end{aligned} \tag{18.15}$$

substitute

$$\begin{aligned} \mathcal{L} &= \frac{1}{2}\sum_{i,j} \lambda_i y_i \vec{x}_i \cdot \vec{x}_j y_j \lambda_j - \sum_{i,j} \lambda_i y_i \vec{x}_i \cdot \vec{x}_j y_j \lambda_j - b\sum_i \lambda_i y_i + \sum_i \lambda_i \\ &= \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i y_i \vec{x}_i \cdot \vec{x}_j y_j \lambda_j \\ &= \sum_i \lambda_i - \frac{1}{2}\sum_{i,j} \lambda_i y_i G_{ij} y_j \lambda_j \end{aligned} \tag{18.16}$$

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

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

$$\sum_i \lambda_i y_i = 0 \tag{18.17}$$

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

$$\lambda_i \left[y_i \left(\vec{w} \cdot \vec{x} + b \right) - 1 \right] = 0 \tag{18.18}$$

$$\begin{aligned} f(\vec{x}) &= \vec{w} \cdot \vec{x} + b \\ &= \sum_i \lambda_i y_i \vec{x}_i \cdot \vec{x} + b \end{aligned} \tag{18.19}$$

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

$$f(\vec{x}) = \vec{W} \cdot \vec{\varphi}(\vec{x}) + b \tag{18.20}$$

$$\mathcal{L} = \sum_i \lambda_i - \sum_{i,j} \lambda_i y_i \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_i) y_j \lambda_j \tag{18.21}$$

nonlinear classifier with kernel trick

$$\mathcal{L} = \sum_i \lambda_i - \sum_{i,j} \lambda_i y_i K_{ij} y_j \lambda_j \tag{18.22}$$

soft margin

$$y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) \geq 1 - s_i \tag{18.23}$$

slack s_i allows violation

$$\begin{aligned} \min \frac{1}{2} \vec{W} \cdot \vec{W} + \lambda_s \sum_i s_i \\ \text{subject to } \; y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) \geq 1 - s_i \\ s_i \geq 0 \end{aligned} \tag{18.24}$$

Lagrangian

$$\mathcal{L} = \frac{1}{2} \vec{W} \cdot \vec{W} + \lambda_s \sum_i s_i - \sum_i \lambda_i \left[y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) - 1 + s_i \right] - \sum_i \lambda_{s,i} s_i \tag{18.25}$$

$$\begin{aligned} 0 &= \frac{\partial \mathcal{L}}{\partial s_i} \\ &= \lambda_s - \lambda_i - \lambda_{s,i} \\ \Rightarrow \lambda_s &= \lambda_i + \lambda_{s,i} \end{aligned} \tag{18.26}$$

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

$$\lambda_i \left[y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) - 1 + s_i \right] = 0 \tag{18.27}$$

$$\lambda_{s,i} s_i = 0 \tag{18.28}$$

$$(\lambda_s - \lambda_i)\, s_i = 0 \tag{18.29}$$

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

$$\begin{aligned} \min \frac{1}{2} \vec{W} \cdot \vec{W} + b^2 + \lambda_s \sum_i s_i \\ \text{subject to } \; y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) \geq 1 - s_i \\ s_i \geq 0 \end{aligned} \tag{18.30}$$

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

$$\mathcal{L} = \frac{1}{2} \vec{W} \cdot \vec{W} + b^2 + \lambda_s \sum_i s_i - \sum_i \lambda_i \left[y_i \left(\vec{W} \cdot \vec{\varphi}(\vec{x}_i) + b \right) - 1 + s_i \right] - \sum_i \lambda_{s,i} s_i \tag{18.31}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \;\Rightarrow\; b = \frac{1}{2} \sum_i \lambda_i y_i \tag{18.32}$$

$$\mathcal{L} = \sum_i \lambda_i - \sum_{i,j} \lambda_i y_i (K_{ij} + 1) y_j \lambda_j \tag{18.33}$$

no equality constraint
solve by SOR relaxation by matrix splitting
least-squares
[Suykens *et al.*, 2002]

$$\min \frac{1}{2}\vec{W}\cdot\vec{W} + \lambda_s \frac{1}{2}\sum_i s_i^2$$
$$\text{subject to } \; y_i\left(\vec{W}\cdot\vec{\varphi}(\vec{x}_i) + b\right) = 1 - s_i \tag{18.34}$$

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

$$\frac{\partial \mathcal{L}}{\partial s_i} = 0 \Rightarrow \lambda_i = \lambda_s s_i \tag{18.35}$$

linear KKT

$$\begin{bmatrix} 0 & y_1 & \cdots & y_N \\ y_1 & & \vdots & \\ \vdots & \cdots & y_i K_{ij} y_j + 1/\lambda_s & \cdots \\ y_N & & \vdots & \end{bmatrix} \begin{bmatrix} b \\ \lambda_1 \\ \vdots \\ \lambda_N \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{18.36}$$

lose sparsity
multiple classes [Vapnik, 1998]
class k

$$f_k(\vec{x}) = \vec{W}_k\cdot\vec{\varphi}(\vec{x}) + b_k \tag{18.37}$$

$$\text{argmax}\{\vec{W}_1\cdot\vec{\varphi}(\vec{x}) + b_1, \ldots, \vec{W}_k\cdot\vec{\varphi}(\vec{x}) + b_k, \ldots\} \tag{18.38}$$

could construct each separately
add constraint and find jointly

$$\vec{W}_k\cdot\vec{\varphi}(\vec{x}_k) + b_k - \vec{W}_m\cdot\vec{\varphi}(\vec{x}_k) + b_m \geq 1 - s_{k,i} \tag{18.39}$$

$m \neq k$

18.3.2 Regression

regression [Smola & Schölkopf, 2004]

$$f(\vec{x}) = \vec{W}\cdot\vec{\varphi}(\vec{x}) + b \tag{18.40}$$

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

$$\begin{aligned} \min \frac{1}{2}\vec{W}\cdot\vec{W} + \lambda_s \sum_i (s_i^+ + s_i^-) \\ \text{subject to} \quad y_i - \vec{W}\cdot\vec{\varphi}(\vec{x}_i) - b \geq \epsilon + s_i^+ \\ \vec{W}\cdot\vec{\varphi}(\vec{x}_i) + b - y_i \geq \epsilon + s_i^- \\ s_i^+, s_i^- \geq 0 \end{aligned} \tag{18.41}$$

Lagrangian

$$\begin{aligned} \mathcal{L} = \frac{1}{2}\vec{W}\cdot\vec{W} + \lambda_s \sum_i \left(s_i^+ + s_i^-\right) - \sum_i \left(\lambda_{s,i}^+ s_i^+ + \lambda_{s,i}^- s_i^-\right) \\ - \sum_i \lambda_i^+ \left(\epsilon + s_i^+ - y_i + \vec{W}\cdot\vec{\varphi}(\vec{x}_i) + b\right) \\ - \sum_i \lambda_i^- \left(\epsilon + s_i^- + y_i - \vec{W}\cdot\vec{\varphi}(\vec{x}_i) - b\right) \end{aligned} \tag{18.42}$$

$$\frac{\partial \mathcal{L}}{\partial \vec{W}} = 0 \;\Rightarrow\; \vec{W} = \sum_i \left(\lambda_i^+ - \lambda_i^-\right)\vec{\varphi}(\vec{x}_i) \tag{18.43}$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \;\Rightarrow\; \sum_i \left(\lambda_i^+ - \lambda_i^-\right) = 0 \tag{18.44}$$

$$\frac{\partial \mathcal{L}}{\partial s_i^+} = 0 \;\Rightarrow\; \lambda_s = \lambda_{s,i}^+ + \lambda_i^+ \tag{18.45}$$

$$\frac{\partial \mathcal{L}}{\partial s_i^-} = 0 \;\Rightarrow\; \lambda_s = \lambda_{s,i}^- + \lambda_i^- \tag{18.46}$$

dual

$$\begin{aligned} \max \; -\frac{1}{2}\sum_{i,j} \left(\lambda_i^+ - \lambda_i^-\right)\left(\lambda_j^+ - \lambda_i^-\right) K_{ij} \\ -\epsilon \sum_i \left(\lambda_i^+ + \lambda_i^-\right) + \sum_i y_i \left(\lambda_i^+ - \lambda_i^-\right) \\ \text{subject to} \quad \sum_i \left(\lambda_i^+ - \lambda_i^-\right) = 0 \\ 0 \leq \lambda_i^+, \lambda_i^- \leq \lambda_s \end{aligned} \tag{18.47}$$

$$\begin{aligned} f(\vec{x}) &= \vec{W}\cdot\vec{\varphi}(\vec{x}) + b \\ &= \sum_i \left(\lambda_i^+ - \lambda_i^-\right) K(\vec{x}, \vec{x}_i) + b \end{aligned} \tag{18.48}$$

18.3.3 Clustering

unsupervised learning

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

$$|\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0|^2 \leq R^2 \tag{18.49}$$

$$|\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0|^2 \leq R^2 + s_i \tag{18.50}$$

want smallest hypersphere: minimize magnitude R

$$\mathcal{L} = R^2 + \lambda_s \sum_i s_i - \sum_i \lambda_i \left[R^2 + s_i - |\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0|^2\right] - \sum_i \lambda_{s,i} s_i \tag{18.51}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial R} &= 2R - 2R \sum_i \lambda_i = 0 \\ &\Rightarrow \sum_i \lambda_i = 1 \end{aligned} \tag{18.52}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \vec{\varphi}_0} &= \frac{\partial}{\partial \vec{\varphi}_0} \sum_i \lambda_i \left(\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0\right) \cdot \left(\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0\right) \\ &= \sum_i \lambda_i \left(-2\vec{\varphi}(\vec{x}_i) + 2\vec{\varphi}_0\right) \\ &= 0 \\ \Rightarrow \vec{\varphi}_0 &= \sum_i \lambda_i \vec{\varphi}(\vec{x}_i) \end{aligned} \tag{18.53}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial s_i} &= \lambda_s - \lambda_i - \lambda_{s,i} = 0 \\ \Rightarrow \lambda_{s,i} &= \lambda_s - \lambda_i \end{aligned} \tag{18.54}$$

KKT

$$\lambda_{s,i} s_i = 0 \tag{18.55}$$

$$\lambda_i \left[R^2 + s_i - |\vec{\varphi}(\vec{x}_i) - \vec{\varphi}_0|^2\right] = 0 \tag{18.56}$$

support vectors

substitute

$$\mathcal{L} = \sum_i \lambda_i \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_i) - \sum_{i,j} \lambda_i \lambda_j \vec{\varphi}(\vec{x}_i) \cdot \vec{\varphi}(\vec{x}_j) \tag{18.57}$$

$$\mathcal{L} = \sum_i \lambda_i K_{ii} - \sum_{i,j} \lambda_i \lambda_j K_{ij} \tag{18.58}$$

$$0 \leq \lambda_i \leq \lambda_s \tag{18.59}$$

radius

$$R(\vec{x}_i) = K_{ii} - 2\sum_j \lambda_j K_{ij} + \sum_{i,j} \lambda_i \lambda_j K_{ij} \tag{18.60}$$

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

18.4 RELAXATIONS

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]
Doyle fragility/complexity

18.5 SELECTED REFERENCES

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

Authoritative survey of convex problems and solutions.

[Schölkopf & Smola, 2002] Schölkopf, B., & Smola, A. J. (2002). *Learning with Kernels*. Cambridge, MA: MIT Press.

A clear tour through kernel methods.

[Cristianini & Shawe-Taylor, 2000] Cristianini, Nello, & Shawe-Taylor, John. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. New York: Cambridge University Press.

A concise introduction to a high-dimensional subject.

18.6 PROBLEMS

(18.1) compressed sensing
choose random frequencies and amplitudes
generate time series
sample random subset of points
equality constraint $\mathbf{A} \cdot \vec{x} - \vec{b} = 0$
calculate minimum L2 norm $\vec{x}$ from SVD
calculate minimum L1 norm $\vec{x}$
approximate L1 norm, minimize exact penalty, increase
compare time series
compare Nyquist requirement

(18.2) SVM