

Mathematics, Optimization, Statistics, and Machine Learning

Sunghee Yun

January 20, 2020

Contents

I	Mathematics	7
1	Calculus	9
1.1	Basics	10
1.2	Multivariate functions	10
1.3	Chain rule	11
1.4	Integration	11
2	Convex analysis	13
2.1	Convex function	14
2.1.1	First order condition	14
2.1.2	Second order condition	17
3	Linear Algebra	19
3.1	Vector space	20
3.2	Eigenvalues	20
3.2.1	Basic definitions	20
3.2.2	Symmetric matrices	21
3.3	Positive definiteness	21
3.4	Matrix norms	22
II	Optimization	23
4	Convex Optimization	25
4.1	Mathematical optimization problem	26
4.2	Convex optimization problem	27
4.3	Duality	28
4.3.1	The Lagrange dual problem	28
4.3.1.1	Examples	28
4.3.2	Interpretations	29
4.3.2.1	Max-min characterization of weak and strong duality	29
4.3.2.2	Saddle-point interpretation	30
4.4	Convex optimization problems	31
4.4.1	Equality constrained problem	31

4.4.1.1	Equality constrained problem examples	31
4.5	Unconstrained minimization	32
4.5.1	Gradient descent method	32
4.5.1.1	Examples	32
5	Portfolio optimization	35
5.1	Problem formulation	34
5.1.1	A portfolio optimization problem	34
III	Statistics	37
6	Statistics Basics	39
6.1	Correlation coefficients	39
6.2	Transformation of a random variable via a function	39
6.2.1	Scale random variable	39
6.2.2	Multivariate random variable	41
6.2.3	Data Examples	41
7	Various distributions	43
7.1	Log-normal distribution	43
7.1.1	Some statistics	44
7.1.2	Parameter estimation	46
8	Bayesian Statistics	47
8.1	Bayesian Theorem	47
8.2	Bayesian Inference	47
8.3	Conjugate prior	47
8.3.1	Bernoulli distribution	47
8.3.2	Gaussian distribution	48
9	Information Theory	49
9.1	Basics	49
9.1.1	Entropy	49
9.1.2	Mutual Information	49
9.1.3	Relative Entropy (Kullback–Leibler divergence)	49
IV	Machine Learning	51
10	Machine Learning Basics	53
10.1	Optimal Predictor	54
10.2	Bias and Variance	54
10.3	Maximum Likelihood Estimation	56
10.4	Maximum a Posteriori Estimation	58
10.5	Bayesian prior update	59
10.6	Predictive Distribution	59

11 Optimization for Machine Learning	61
11.1 Gradient method	62
11.2 Stochastic gradient method	62
12 Bayesian Network	63
13 Collaborative Filtering	65
13.1 Item-based Collaborative Filtering	66
13.1.1 Rating matrix modeling for menu personalization for mobile shopping app	66
13.1.1.1 Rating matrix modeling	67
13.1.2 Value augmentation based on Bayesian MAP	68
13.1.3 Similarity measure among items	68
13.1.3.1 Cosine similarity	69
13.1.3.2 Cosine similarity when prior distribution is used	69
13.1.3.3 Correlation coefficient similarity	70
13.1.4 Data value transformation	70
13.1.4.1 TFIDF (or tf-idf)	71
13.1.4.2 Okapi BM25 transformation	71
13.1.5 Recommendation based on item similarities	71
13.2 Collaborative Filtering using Matrix Factorization	72
13.2.1 Problem definition and formulations	72
13.2.2 Solution methods	74
13.2.2.1 Matrix factorization via singular value decomposition (SVD)	74
13.2.2.2 Matrix factorization via gradient descent (GD) method	75
13.2.2.3 Matrix factorization via alternating gradient descent (GD) method	77
13.2.2.4 Matrix factorization via stochastic gradient descent (SGD) method	77
13.2.2.5 Matrix factorization via alternating least-squares (ALS)	78
13.2.2.6 Weighted matrix factorization via alternating least-squares (ALS)	80
13.2.3 Collaborative filtering for implicit feedback dataset	80
13.2.3.1 Regularization coefficient conversion	81
13.3 Appendix	83
13.3.1 Linear algebra	83
13.3.1.1 Singular value decomposition (SVD)	83
13.3.1.2 Singular value decomposition as rank- k approximation	84
14 Time Series Anomaly Detection	85
14.1 Real-Time Anomaly Detection	86
14.1.1 Computing Anomaly Likelihood	86
15 Reinforcement Learning	89
15.1 Finite Markov decision processes	90
15.1.1 Markov property	90
15.1.2 Policy	91
15.1.3 Return	91
15.1.4 State value function and action value function	92
15.2 Bellman equation	92
15.2.1 Bellman equations	92

15.2.2	Bellman optimality equations	93
15.3	Dynamic programming	94
15.3.1	Policy evaluation (prediction)	94
15.3.2	Policy iteration	95
15.3.3	Value iteration	95
15.4	Monte Carlo methods	95
15.4.1	Monte Carlo prediction	97
15.4.2	Monte Carlo control	98
15.4.3	Monte Carlo control without exploring starts	98
15.4.4	Off-policy prediction via important sampling	100
15.4.5	Off-policy Monte Carlo control	100
15.5	Temporal-difference learning	100
15.5.1	TD prediction	100
15.5.2	Sarsa: on-policy TD Control	102
15.5.3	Q-learning: off-policy TD control	102
15.5.4	Maximization bias and double learning	103
15.6	n -step bootstrapping	103
15.6.1	n -step TD prediction	104
15.6.2	n -step Sarsa	108
15.6.3	n -step off-policy learning	110
15.7	Planning and learning with tabular methods	110
15.7.1	Dyna: integrated planning, acting, and learning	110
15.8	On-policy Prediction with Approximation	110
15.9	On-policy Control with Approximation	110
15.10	Off-policy Methods with Approximation	113
15.11	Eligibility Traces	113
15.11.1	The λ -return	113
15.11.2	$TD(\lambda)$	114
15.11.3	Why $TD(\lambda)$ approximates the off-line λ -return algorithm?	115
15.11.4	Sarsa(λ)	120
15.11.5	Tabular methods using eligibility traces	120
15.12	Appendix: conditional probability and expected value	122

Part I

Mathematics

Chapter 1

Calculus

Chapter 2

Convex analysis

Chapter 3

Linear Algebra

Part II

Optimization

Chapter 4

Convex Optimization

4.1 Mathematical optimization problem

A mathematical optimization problem can be expressed as

$$\begin{aligned} & \text{minimize} && f_0(x) \\ & \text{subject to} && f_i(x) \leq 0 \text{ for } i = 1, \dots, m \\ & && h_i(x) = 0 \text{ for } i = 1, \dots, p \end{aligned} \quad (4.1)$$

where $x \in \mathbf{R}^n$ is the optimization variable, $f_0 : \mathbf{R}^n \rightarrow \mathbf{R}$ is the objective function, $f_i : \mathbf{R}^n \rightarrow \mathbf{R}$ for $i = 1, \dots, m$ are the inequality constraint functions, and $h_i : \mathbf{R}^n \rightarrow \mathbf{R}$ for $i = 1, \dots, p$ are the equality constraint functions.

The conditions, $f_i(x) \leq 0$ for $i = 1, \dots, m$, are called inequality constraints and the conditions, $h_i(x) = 0$ for $i = 1, \dots, p$ are called equation constraints.

Note that this formulation covers pretty much every single-objective optimization problem. For example, consider the following optimization problem.

$$\begin{aligned} & \text{maximize} && f(x_1, x_2) \\ & \text{subject to} && x_1 \geq x_2 \\ & && x_1 + x_2 = 2 \end{aligned} \quad (4.2)$$

This problem can be cast into an equivalent problem as follows.

$$\begin{aligned} & \text{minimize} && -f(x_1, x_2) \\ & \text{subject to} && -x_1 + x_2 \leq 0 \\ & && x_1 + x_2 - 2 = 0 \end{aligned} \quad (4.3)$$

The feasible set for (4.1) is defined by the set of $x \in \mathbf{R}^n$ which satisfies all the constraints. Also, the optimal value for (4.1) is the infimum of $f_0(x)$ while x is in the feasible set. When the infimum is achievable, we define the optimal solution set as the set of all feasible x achieving the infimum value. These are defined in mathematically rigorous terms below.

- The feasible set for (4.1) is defined by

$$\mathcal{F} = \{x \in \mathcal{D} \mid f_i(x) \leq 0 \text{ for } i = 0, \dots, m, h_j(x) = 0 \text{ for } j = 1, \dots, p\} \subseteq \mathbf{R}^n \quad (4.4)$$

where

$$\mathcal{D} = \left(\bigcap_{0 \leq i \leq m} \text{dom } f_i \right) \cap \left(\bigcap_{1 \leq i \leq p} \text{dom } h_i \right). \quad (4.5)$$

- The optimal value for (4.1) is defined by

$$p^* = \inf_{x \in \mathcal{F}} f_0(x) \quad (4.6)$$

We use the conventions that $p^* = -\infty$ if $f_0(x)$ is unbounded below for $x \in \mathcal{F}$ and that $p^* = \infty$ if $\mathcal{F} = \emptyset$.

- The optimal solution set for (4.1) is defined by

$$\mathcal{X}^* = \{x \in \mathcal{F} \mid f_0(x) = p^*\}. \quad (4.7)$$

4.2 Convex optimization problem

A mathematical optimization problem is called a convex optimization problem if the objective function and all the inequality constraint functions are convex functions and all the equality constraint functions are affine functions.

Hence, a convex optimization problem can be expressed as

$$\begin{aligned} & \text{minimize} && f_0(x) \\ & \text{subject to} && f_i(x) \leq 0 \text{ for } i = 1, \dots, m \\ & && Ax = b \end{aligned} \tag{4.8}$$

where $x \in \mathbf{R}^n$ is the optimization variable, $f_i : \mathbf{R}^n \rightarrow \mathbf{R}$ for $i = 0, \dots, n$ are convex functions, $h_i : \mathbf{R}^n \rightarrow \mathbf{R}$ for $i = 1, \dots, p$ are the equality constraint functions, $A \in \mathbf{R}^{p \times n}$, and $b \in \mathbf{R}^p$.

A function, $f : \mathbf{R}^n \rightarrow \mathbf{R}$, is called a convex function if $\text{dom } f \subseteq \mathbf{R}^n$ is a convex set and for all $x, y \in \text{dom } f$ and all $0 \leq \lambda \leq 1$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \tag{4.9}$$

where $\text{dom } f \subseteq \mathbf{R}^n$ denotes the domain of f .

A convex optimization enjoys a number of nice theoretical and practical properties.

- A local minimum of a convex optimization problem is a global minimum, *i.e.*, if for some $R > 0$ and $x_0 \in \mathcal{F}$, $\|x - x_0\| < R$ and $x \in \mathcal{F}$ imply $f_0(x_0) \leq f_0(x)$, then $f_0(x_0) \leq f_0(x)$ for all $x \in \mathcal{F}$.

Proof: Assume that $x_0 \in \mathcal{F}$ is a local minimum, *i.e.*, for some $R > 0$, $\|x - x_0\| < R$ and $x \in \mathcal{F}$ imply $f_0(x_0) \leq f_0(x)$.

Now assume that x_0 is not a global minimum, *i.e.*, there exists $y \in \mathcal{F}$ such that $y \neq x_0$ and $f_0(y) < f_0(x_0)$. Then for $z = \lambda y + (1 - \lambda)x_0$ with $\lambda = \min\{R/\|y - x_0\|, 1\}/2$, the convexity of f_0 implies

$$f_0(z) \leq \lambda f_0(y) + (1 - \lambda)f_0(x_0) \tag{4.10}$$

since $0 < \lambda \leq 1/2 < 1$. Furthermore

$$\|z - x_0\| = \lambda\|y - x_0\| \leq R/2, \tag{4.11}$$

hence $f_0(z) \geq f_0(x_0)$, which together with (4.10) implies

$$f_0(x_0) \leq f_0(z) \leq \lambda f_0(y) + (1 - \lambda)f_0(x_0) < \lambda f_0(x_0) + (1 - \lambda)f_0(x_0) = f_0(x_0), \tag{4.12}$$

which is a contradiction. Therefore there is no $y \in \mathcal{F}$ such that $y \neq x_0$ and $f_0(y) < f_0(x_0)$. Therefore x_0 is a global minimum.

- For a unconstrained problem, *i.e.*, the problem (4.8) with $m = p = 0$, with differentiable objective function, $x \in \text{dom } f_0$ is an optimal solution if and only if $\nabla f_0(x) = 0 \in \mathbf{R}^n$.

Proof: The Taylor theorem implies that for any $x, y \in \mathbf{dom} f_0$,

$$f_0(y) = f_0(x) + \nabla f_0(x)^T(y - x) + \frac{1}{2}(y - x)^T \nabla^2 f_0(z)(y - x) \quad (4.13)$$

for some z on the line segment having x and y as its end points, *i.e.*, $z = \alpha x + (1 - \alpha)y$ for some $0 \leq \alpha \leq 1$. Since $\nabla^2 f_0(x) \succeq 0$ for any $z \in \mathbf{dom} f_0$, we have

$$f_0(y) \geq f_0(x) + \nabla f_0(x)^T(y - x) \quad (4.14)$$

Thus, if for some $x_0 \in \mathbf{R}^n$, $\nabla f_0(x_0) = 0$, for any $x \in \mathbf{dom} f_0$,

$$f_0(x) \geq f_0(x_0) + \nabla f_0(x_0)^T(x - x_0) = f_0(x_0), \quad (4.15)$$

hence x_0 is an optimal solution. Now assume that x_0 is an optimal solution, but $\nabla f_0(x_0) \neq 0$. Then for any $k > 0$, if we let $x = x_0$ and $y = x_0 - k \nabla f_0(x_0)$, (4.13) becomes

$$\begin{aligned} f_0(y) &= f_0(x_0) + \nabla f_0(x_0)^T(-k \nabla f_0(x_0)) + \frac{k^2}{2} \nabla f_0(x_0)^T \nabla^2 f_0(z) \nabla f_0(x_0) \\ &= f_0(x_0) - k \|\nabla f_0(x_0)\|^2 + \frac{k^2}{2} \nabla f_0(x_0)^T \nabla^2 f_0(z) \nabla f_0(x_0) \end{aligned}$$

for all $y = x_0 - k \nabla f_0(x_0) \in \mathbf{dom} f_0$.

Since for $k < 2\|\nabla f_0(x_0)\|^2 / \nabla f_0(x_0)^T \nabla^2 f_0(z) \nabla f_0(x_0)$, $-k\|\nabla f_0(x_0)\|^2 + \frac{k^2}{2} \nabla f_0(x_0)^T \nabla^2 f_0(z) \nabla f_0(x_0) < 0$, thus $f_0(y) < f_0(x_0)$, hence the contradiction. Therefore, if x_0 is an optimal solution for the unconstrained problem, $\nabla f_0(x_0) = 0$.

4.3 Duality

4.3.1 The Lagrange dual problem

4.3.1.1 Examples

4.3.1.1.1 Standard form LP

$$\begin{aligned} &\text{minimize} && c^T x \\ &\text{subject to} && Ax = b \\ &&& x \succeq 0 \end{aligned} \quad (4.16)$$

The Lagrange dual problem is

$$\begin{aligned} &\text{maximize} && -b^T \nu \\ &\text{subject to} && A^T \nu + c \geq 0 \end{aligned} \quad (4.17)$$

4.3.1.1.2 Inequality form LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \end{array} \quad (4.18)$$

The Lagrange dual problem is

$$\begin{array}{ll} \text{maximize} & -b^T \lambda \\ \text{subject to} & A^T \lambda + c = 0 \\ & \lambda \succeq 0 \end{array} \quad (4.19)$$

4.3.1.1.3 Least-squares solution of linear equations

$$\begin{array}{ll} \text{minimize} & (1/2)x^T x \\ \text{subject to} & Ax = b \end{array} \quad (4.20)$$

The Lagrange dual problem is

$$\text{maximize} \quad -(1/2)\nu^T A A^T \nu - b^T \nu \quad (4.21)$$

4.3.1.1.4 Entropy maximization

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n x_i \log x_i \\ \text{subject to} & Ax = b \\ & \mathbf{1}^T x = 1 \end{array} \quad (4.22)$$

with domain $\mathcal{D} = \mathbf{R}_+^n$

The Lagrange dual problem is

$$\begin{array}{ll} \text{maximize} & -b^T \lambda - \log \left(\sum_{i=1}^n \exp(-a_i^T \lambda) \right) \\ \text{subject to} & \lambda \succeq 0 \end{array} \quad (4.23)$$

4.3.2 Interpretations**4.3.2.1 Max-min characterization of weak and strong duality**

We first note that for any $f : X \times Y \rightarrow \mathbf{R}$, we have

$$\sup_{y \in Y} \inf_{x \in X} f(x, y) \leq \inf_{x \in X} \sup_{y \in Y} f(x, y). \quad (4.24)$$

This inequality is called *max-min inequality*.

We can prove this as follows. Let $g : Y \rightarrow \mathbf{R}$ be a function defined by $g(y) = \inf_{x \in X} f(x, y)$ and let $h : X \rightarrow \mathbf{R}$ be a function defined by $h(x) = \sup_{y \in Y} f(x, y)$. Then we have that for any $x \in X$ and $y \in Y$

$$g(y) = \inf_{x \in X} f(x, y) \leq f(x, y), \quad (4.25)$$

which implies that for any $x \in X$

$$\sup_{y \in Y} g(y) \leq \sup_{y \in Y} f(x, y) = h(x). \quad (4.26)$$

This again implies that

$$\sup_{y \in Y} g(y) \leq \inf_{x \in X} h(x), \quad (4.27)$$

hence the proof.

4.3.2.2 Saddle-point interpretation

Suppose $f : X \times Y \rightarrow \mathbf{R}$. We refer a point $(\tilde{x}, \tilde{y}) \in X \times Y$ a *saddle-point* for f (and X and Y) if

$$f(\tilde{x}, y) \leq f(\tilde{x}, \tilde{y}) \leq f(x, \tilde{y}) \quad (4.28)$$

for all $x \in X$ and $y \in Y$.

Now if x^* and λ^* are primal and dual optimal points for a problem in which strong duality obtains, the form a saddle-point for the Lagrangian. Conversely, if (x, λ) is a saddle-point of the Lagrangian, then x is primal optimal, λ is dual optimal, and the optimal duality gap is zero.

To prove these, assume that $x^* \in \mathcal{D}$ and $(\lambda^*, \nu^*) \in \mathbf{R}_+^m \times \mathbf{R}^p$ are primal and dual optimal points for a problem in which strong duality obtains. Then for any $x \in \mathcal{D}$ and $(\lambda, \nu) \in \mathbf{R}_+^m \times \mathbf{R}^p$, we have

$$L(x^*, \lambda, \nu) = f_0(x^*) + \sum_{i=1}^m \lambda_i f_i(x^*) + \sum_{i=1}^p \nu_i h_i(x^*) \leq f_0(x^*) = g(\lambda^*, \nu^*) \leq L(x, \lambda^*, \nu^*) \quad (4.29)$$

where the left inequality comes from the fact that $\lambda_i f_i(x^*) \leq 0$ for all $i = 1, \dots, m$ and $h_i(x^*) = 0$ for all $i = 1, \dots, p$ and the right inequality comes from the definition of (Lagrange) dual function. Now from the complementary slackness we know that $\lambda_i f_i(x^*) = 0$ for all $i = 1, \dots, m$. Therefore

$$L(x^*, \lambda^*, \nu^*) = f_0(x^*), \quad (4.30)$$

thus we have

$$L(x^*, \lambda, \nu) \leq L(x^*, \lambda^*, \nu^*) \leq L(x, \lambda^*, \nu^*), \quad (4.31)$$

hence the proof.

Now suppose that $\tilde{x} \in \mathcal{D}$ and $(\tilde{\lambda}, \tilde{\nu}) \in \mathbf{R}_+^m \times \mathbf{R}^p$ are the saddle-point of the Lagrangian, *i.e.*, for all $x \in \mathcal{D}$ and $(\lambda, \nu) \in \mathbf{R}_+^m \times \mathbf{R}^p$,

$$L(\tilde{x}, \lambda, \nu) \leq L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \leq L(x, \tilde{\lambda}, \tilde{\nu}). \quad (4.32)$$

First we show that \tilde{x} is a feasible point. The left inequality says that for all $(\lambda, \nu) \in \mathbf{R}_+^m \times \mathbf{R}^p$,

$$L(\tilde{x}, \lambda, \nu) = f_0(\tilde{x}) + \sum_{i=1}^m \lambda_i f_i(\tilde{x}) + \sum_{i=1}^p \nu_i h_i(\tilde{x}) \leq L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \quad (4.33)$$

If $f_i(\tilde{x}) > 0$ for some $i \in \{1, \dots, m\}$ or $h_i(\tilde{x}) \neq 0$ for some $i \in \{1, \dots, p\}$, $L(\tilde{x}, \lambda, \nu)$ is unbounded above and the above inequality cannot hold. Therefore $f_i(\tilde{x}) \leq 0$ for all $i \in \{1, \dots, m\}$ and $h_i(\tilde{x}) = 0$

for all $i \in \{1, \dots, p\}$, *i.e.*, \tilde{x} is primal feasible. Since the inequality must hold when $\lambda = 0$ and $\nu = 0$, we have

$$f(\tilde{x}) \leq L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}). \quad (4.34)$$

The right inequality of (4.32) implies that

$$L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \leq g(\tilde{\lambda}, \tilde{\nu}) = \inf_{x \in \mathcal{D}} L(x, \tilde{\lambda}, \tilde{\nu}), \quad (4.35)$$

which implies that $f_0(\tilde{x}) \leq g(\tilde{\lambda}, \tilde{\nu})$. Since $g(\lambda, \nu)$ is an underestimator of $f_0(x)$ for any feasible $x \in \mathcal{D}$ and $(\tilde{\lambda}, \tilde{\nu}) \in \mathbf{R}_+^m \times \mathbf{R}^p$, *i.e.*, $g(\tilde{\lambda}, \tilde{\nu}) \leq f_0(\tilde{x})$, thus $g(\tilde{\lambda}, \tilde{\nu}) = f_0(\tilde{x})$. Therefore \tilde{x} is an optimal solution for the primal problem and $(\tilde{\lambda}, \tilde{\nu})$ is an optimal solution for the dual problem, hence the proof.

4.4 Convex optimization problems

4.4.1 Equality constrained problem

Consider the following equality constrained problem:

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && Ax = b \end{aligned} \quad (4.36)$$

where $x \in \mathbf{R}^n$ is the optimization variable, $A \in \mathbf{R}^{m \times n}$, and $b \in \mathbf{R}^m$. The Lagrangian is

$$L(x, \nu) = f(x) + \nu^T (Ax - b) \quad (4.37)$$

and the Lagrange dual function is

$$g(\nu) = \inf_{x \in \mathbf{R}^n} L(x, \nu) = - \sup_{x \in \mathbf{R}^n} (-\nu^T Ax - f(x)) - b^T \nu = -f^*(-A^T \nu) - b^T \nu \quad (4.38)$$

The KKT optimality conditions are

$$\text{primal feasibility:} \quad Ax = b \quad (4.39)$$

$$\text{gradient of Lagrangian vanishes:} \quad \nabla f(x) + A^T \nu = 0 \quad (4.40)$$

4.4.1.1 Equality constrained problem examples

Consider the following equality constraint quadratic problem:

$$\begin{aligned} & \text{minimize} && x^T P x + q^T x \\ & \text{subject to} && Ax = b \end{aligned} \quad (4.41)$$

where $x \in \mathbf{R}^n$ is the optimization variable, $P \in \mathcal{S}_{++}^n$, $q \in \mathbf{R}^n$, $A \in \mathbf{R}^{m \times n}$, and $b \in \mathbf{R}^m$.

The Lagrangian is

$$L(x, \nu) = x^T P x + q^T x + \nu^T (Ax - b). \quad (4.42)$$

The gradient of the Lagrangian with respect to x is

$$\nabla_x L(x, \nu) = 2Px + q + A^T \nu = 0, \quad (4.43)$$

hence

$$\operatorname{argmin}_x L(x, \nu) = -\frac{1}{2}P^{-1}(q + A^T \nu) \quad (4.44)$$

The KKT conditions are

$$\text{primal feasibility:} \quad Ax = b \quad (4.45)$$

$$\text{gradient of Lagrangian vanishes:} \quad 2Px + q + A^T \nu = 0 \quad (4.46)$$

which are equivalent to

$$\begin{bmatrix} 2P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \nu \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}. \quad (4.47)$$

The conjugate of the objective function is

$$f^*(y) = \sup_x (y^T x - x^T P x - q^T x). \quad (4.48)$$

Since the gradient of $y^T x - x^T P x - q^T x$ is $y - q - 2Px$,

$$\operatorname{argsup}_x (y^T x - x^T P x - q^T x) = \frac{1}{2}P^{-1}(y - q), \quad (4.49)$$

thus

$$\begin{aligned} f^*(y) &= -\frac{1}{4}(y - q)^T P^{-1}(y - q) + \frac{1}{2}(y - q)^T P^{-1}(y - q) = \frac{1}{4}(y - q)^T P^{-1}(y - q) \\ &= \frac{1}{4}(y^T P^{-1}y - 2q^T P^{-1}y + q^T P^{-1}q) \end{aligned}$$

4.5 Unconstrained minimization

4.5.1 Gradient descent method

4.5.1.1 Examples

4.5.1.1.1 A quadratic problem in \mathbf{R}^2 We consider the quadratic objective function on \mathbf{R}^2

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2) \quad (4.50)$$

where $\gamma > 0$.

We apply the gradient descent method with exact line search. The gradient of f is

$$\nabla f(x) = \begin{bmatrix} x_1 \\ \gamma x_2 \end{bmatrix} \quad (4.51)$$

Let $\tilde{f} : \mathbf{R}_+ \rightarrow \mathbf{R}$ defined by $\tilde{f}(t) = f(x - t\nabla f(x))$. Now

$$\tilde{f}(t) = f\left(\begin{bmatrix} (1-t)x_1 \\ (1-\gamma t)x_2 \end{bmatrix}\right) = \frac{1}{2}((1-t)^2x_1^2 + \gamma(1-\gamma t)^2x_2^2) \quad (4.52)$$

and

$$\frac{d}{dt}\tilde{f}(t) = -(1-t)x_1^2 - \gamma^2(1-\gamma t)x_2^2 = 0 \quad (4.53)$$

implies

$$t = \frac{x_1^2 + \gamma^2x_2^2}{x_1^2 + \gamma^3x_2^2} \quad (4.54)$$

minimizes $\tilde{f}(t)$. Since

$$1-t = \frac{\gamma^2(\gamma-1)x_2^2}{x_1^2 + \gamma^3x_2^2} \quad (4.55)$$

and

$$1-\gamma t = \frac{(1-\gamma)x_1^2}{x_1^2 + \gamma^3x_2^2} \quad (4.56)$$

Thus the exact line search yields

$$x^+ = x - t\nabla f(x) = \begin{bmatrix} (1-t)x_1 \\ (1-\gamma t)x_2 \end{bmatrix} = \frac{(1-\gamma)x_1x_2}{x_1^2 + \gamma^3x_2^2} \begin{bmatrix} -\gamma^2x_2 \\ x_1 \end{bmatrix}. \quad (4.57)$$

If $x = \alpha[\gamma \ 1]^T$, then

$$x^+ = \frac{\alpha^3(1-\gamma)\gamma}{\alpha^2\gamma^2(1+\gamma)} \begin{bmatrix} -\gamma^2 \\ \gamma \end{bmatrix} = \alpha \frac{1-\gamma}{1+\gamma} \begin{bmatrix} -\gamma \\ 1 \end{bmatrix}. \quad (4.58)$$

If $x = \alpha[-\gamma \ 1]^T$, then

$$x^+ = -\frac{\alpha^3(1-\gamma)\gamma}{\alpha^2\gamma^2(1+\gamma)} \begin{bmatrix} -\gamma^2 \\ -\gamma \end{bmatrix} = \alpha \frac{1-\gamma}{1+\gamma} \begin{bmatrix} \gamma \\ 1 \end{bmatrix}. \quad (4.59)$$

Therefore if $x^{(0)} = [\gamma \ 1]^T$, then

$$x^{(k)} = \left(\frac{1-\gamma}{1+\gamma}\right)^k \begin{bmatrix} (-1)^k\gamma \\ 1 \end{bmatrix} = \left(\frac{\gamma-1}{\gamma+1}\right)^k \begin{bmatrix} \gamma \\ (-1)^k \end{bmatrix}. \quad (4.60)$$

Chapter 5

Portfolio optimization

Part III

Statistics

Part IV

Machine Learning

Chapter 10

Machine Learning Basics

Chapter 11

Optimization for Machine Learning

Chapter 12

Bayesian Network

Chapter 13

Collaborative Filtering

Chapter 14

Time Series Anomaly Detection

Chapter 15

Reinforcement Learning