MATH 408 Notes

Contents

1 Section One

1.1 Vector Space $\mathbb{R}^n$

$\mathbb{R}^n$ is the set of real column vectors $x = (x_1, ..., x_n)$

- Addition

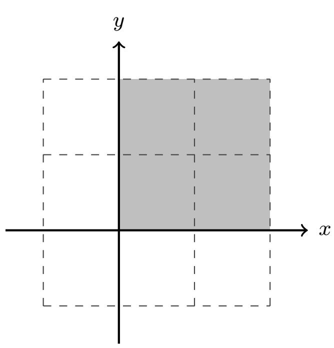
$$x + y = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

- Scalar multiplication

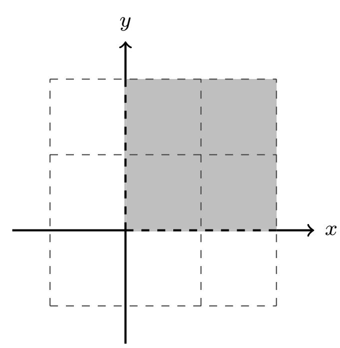
$$\lambda x = \lambda \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{bmatrix}$$

1.2 Subsets of $\mathbb{R}^n$

- $\mathbb{R}^n_+ = \{(x_1, ..., x_n) \mid x_1, ..., x_n \geq 0\}$



- $\mathbb{R}^n_{++} = \{(x_1, ..., x_n) \mid x_1, ..., x_n > 0\}$



1.3 Line Segments

- Closed line segment
  The closed line segment $[x, y]$ between the points x and y is the set $\{\lambda x + (1 - \lambda) y \mid \lambda \in [0, 1]\}$
- Open line segment
  The open line segment (x, y) between the points x and y is the set $\{\lambda x + (1 - \lambda) y \mid \lambda \in (0, 1)\}$

1.4 Unit Simplex

The unit simplex Δ_n is the set $\{x \in \mathbb{R}^n_+ \mid x_1 + ... + x_n = 1\}$

1.5 Polyhedrons

A polyhedron P is the set of points $\{x \mid {a_i}^T x \leq b_i,\ \forall i = 1, ..., k\}$

1.6 Matrix Space $\mathbb{R}^{m\times n}$

$\mathbb{R}^{m\times n}$ is the set of real matrices $A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix}$

- Addition

  $$A + B = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} + \begin{bmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{m1} & \dots & b_{mn} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & \dots & a_{1n}+b_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & \dots & a_{mn}+b_{mn} \end{bmatrix}$$

- Scalar Multiplication

  $$\lambda A = \lambda \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} \lambda a_{11} & \dots & \lambda a_{1n} \\ \vdots & \ddots & \vdots \\ \lambda a_{m1} & \dots & \lambda a_{mn} \end{bmatrix}$$

- Square Matrix Trace

  $$\mathrm{tr}(A) = \mathrm{tr}\left(\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}\right) = a_{11} + a_{22} + ... + a_{nn}$$

1.7 Subsets of $\mathbb{R}^{m\times n}$

- Symmetric matrices

  $S^n = \left\{A \in \mathbb{R}^{n\times n} \mid A = A^T\right\}$

- Positive semidefinite matrices

  $S^n_+ = \left\{A \in S^n \mid x^T A x \geq 0,\ \forall x \in \mathbb{R}^n\right\}$

  – If $A \in S^n_+$, then we can denote this as $A \succeq 0$

- Positive definite matrices

  $S^n_{++} = \left\{A \in S^n \mid x^T A x > 0,\ \forall x \in \mathbb{R}^n \backslash \{0\}\right\}$

  – If $A \in S^n_{++}$, then we can denote this as $A \succ 0$

- Orthogonal matrices

  $\mathbb{O}^n = \left\{A \in \mathbb{R}^{n\times n} \mid A^T A = I\right\}$

1.8 Dot Product

The dot product operation $x^T y$ can be denoted as $\langle x, y\rangle$

- $\langle x, y\rangle = \langle y, x\rangle$
- $\langle x, y+z\rangle = \langle x, y\rangle + \langle x, z\rangle$
- $\langle \lambda x, y\rangle = \lambda \langle x, y\rangle$
- $\langle x, x\rangle \geq 0$ for all x
- $\langle x, x\rangle = 0$ if and only if $x = 0$

1.9 Vector Norms

The norm $||a||$ is a number assigned to each real or complex n-vector a. Vector norms satisfy the following properties

- For all vectors a, $||a|| \geq 0$ and $||a|| = 0$ if and only if $a = 0$
  - The only vector with zero length is the zero vector
- For vectors a and all scalars $\alpha \in \mathbb{R}$ or $\mathbb{C}$, $||\alpha a|| = |\alpha| \cdot ||a||$
  - Scaling a vector also scales its norm
- For all vectors a, b, $||a + b|| \leq ||a|| + ||b||$
  - In a triangle, the sum of lengths of two sides is greater than or equal to the length of the remaining side

Common vector norms

- $||a||_1 = \sum_{j=1}^{n} |a_j|$
  - Referred to as the 'one norm'
  - This is the absolute vector sum
- $||a||_2 = \left(\sum_{j=1}^{n} |a_j|^2 \right)^{\frac{1}{2}}$
  - Referred to as the 'two/Euclidean norm'
  - This is the root of the absolute square vector sum
- $||a||_\infty = \max_{1 \leq j \leq n} |a_j|$
  - Referred to as the 'infinity/max norm'
  - This is the maximum absolute element

1.10 Cauchy-Schwartz Inequality

The Cauchy-Schwartz inequality states that $| \langle x, y \rangle | \leq ||x||_2 \cdot ||y||_2$

- Equality holds if and only if x and y are linearly independent
- $\langle x, y \rangle = ||x||_2 \cdot ||y||_2 \cdot \cos\theta$

1.11 Matrix Norms

The operator norm of an $(n \times n)$ matrix is $||A||_{\text{op}} = \sup_{x : ||x||_2 \leq 1} ||Ax||_2$

- $||Ax||_2 \leq ||A||_{\text{op}} \cdot ||x||_2$

1.12 Frobenius Norm

The Frobenius norm of an $(m \times n)$ matrix is $||A||_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{\frac{1}{2}}$

- $||A||_F = \text{tr}(A^T A)^{\frac{1}{2}}$

1.13 Eigenvalue Decompositions

Let $A \in S^n$. Then a scalar $\lambda \in \mathbb{R}$ is an eigenvalue of A if $A - \lambda I$ is singular

- Any $u \in \text{null}(A - \lambda I)$ where $u \neq 0$ is an eigenvector of A
- If $A \in S^n$ is symmetric, then the polynomial $p(\lambda) = \det(A - \lambda I)$ has exactly n real roots, including multiplicities
- A matrix $A \in S^n$ has at most n eigenvalues
- Given that $A \in S^n$ has eigenvalues $\lambda_1, ..., \lambda_n$
  - $\text{tr(A)} = \lambda_1 + ... + \lambda_n$
  - $\det(A) = \lambda_1 ... \lambda_n$

1.14 Spectral Decomposition Theorem

Let $A \in S^n$. Then there exists $U \in \mathbb{O}^n$ and a diagonal matrix $\Omega = \text{diag}(\lambda_1, ..., \lambda_n)$ satisfying $A = U \Omega U^T$

1.15 Rayleigh-Ritz Theorem

Let $A \in S^n$. Then $\lambda_{\min} ||x||_2{}^2 \leq \langle Ax, x \rangle \leq \lambda_{\max} ||x||_2{}^2$ where $\lambda_{\min}$ is the minimum eigenvalue of A and $\lambda_{\max}$ is the maximum eigenvalue of A

1.16 Balls

An ball in $\mathbb{R}^n$ is the volume of space bounded by an n-dimensional ball

- Open ball

  $B(x, r) = \{y \in \mathbb{R}^n \mid ||y - x||_2 < r\}$
- Closed ball

  $B[x, r] = \{y \in \mathbb{R}^n \mid ||y - x||_2 \leq r\}$

1.17 Interior Points

A point $x \in U$ where $U \subseteq \mathbb{R}^n$ is an interior point of a volume U if there exists $r > 0$ such that $B(x, r) \subseteq U$

- A point $x \in U$ is an interior point of U if there exists a ball with non-zero radius that is fully enclosed within U

1.18 Interiors

The interior of a volume U where $U \subseteq \mathbb{R}^n$ is the set of all interior points of U

- $\text{int}(U) = \{x \in U \mid x \text{ is an interior point}\}$

1.19 Open Sets

A set U is an open set if $U = \mathrm{int}(U)$

- U is an open set if and only if U contains no boundary points
- The union of any number of open sets is open
- The intersection of finitely many open sets is open

1.20 Closed Sets

A set U is a closed set if its complement $U^c = \{x \in \mathbb{R}^n \mid x \notin U\}$ is open

- U is a closed set if and only if every sequence $x_n \in U$ converges to a point in U

1.21 Boundaries

The boundary of a set U is the set of non-interior points of U

- $\mathrm{bd}(U) = \{x \in U \mid B(x,r) \cap U \neq \varnothing \text{ and } B(x,r) \cap U^c \neq \varnothing \text{ for } r > 0\}$

1.22 Closure

The closure of a set U is the union of U and its boundary

- $\mathrm{cl}(U) = U \cup \mathrm{bd}(U)$
- The closure of U is the smallest

1.23 Bounded Sets

A set U is bounded if there exists $r > 0$ such that $U \subseteq B(0,r)$

1.24 Compact Sets

A set U is compact if U is closed and bounded

1.25 Continuous Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous if $\lim_{y \to x} f(y) = f(x)$ for all $x \in \mathbb{R}^n$

- If f is continuous, then the following sets are closed
  - $[f = r] = \{x \in \mathbb{R}^n \mid f(x) = r\}$
  - $[f \leq r] = \{x \in \mathbb{R}^n \mid f(x) \leq r\}$

1.26 Extreme Value Theorem

Any continuous function $f : U \to \mathbb{R}$ defined on a compact set U contains its infimum and supremum

1.27 Minimizer

An element $\bar{x}$ is a minimizer of f if $f(\bar{x}) \leq f(x)$ for all $x \in \mathbb{R}^n$

- If a minimizer $\bar{x}$ exists, then $f(\bar{x}) = \inf f(x)$
- A minimizer $\bar{x}$ does not necessarily exist
  - i.e. $f(x) = \frac{1}{x}$ does not have a minimizer

1.28 First-Order Partial Derivatives

Let $f : U \to \mathbb{R}$ where the set $U \subseteq \mathbb{R}^n$ is open. Then $\dfrac{\partial f}{\partial x_i}(x)$ is the first-order partial derivative of f with respect to x_i

$$\frac{\partial f}{\partial x_i}(x) = \lim_{t \to 0} \frac{f(x + te_i) - f(x)}{t}$$

1.29 Second-Order Partial Derivatives

Let $f : U \to \mathbb{R}$ where the set $U \subseteq \mathbb{R}^n$ is open and let $g(x) = \dfrac{\partial f}{\partial x_i}(x)$. Then $\dfrac{\partial f}{\partial x_j \partial x_i}(x)$ is the second-order partial derivative of f

$$\frac{\partial f}{\partial x_j \partial x_i}(x) = \frac{\partial g}{\partial x_j}(x)$$

1.30 C'-Smooth

A function is C'-smooth if $\dfrac{\partial f}{\partial x_i}(x)$ exists and is continuous for all $i = 1, ..., n$

1.31 C''-Smooth

A function is C''-smooth if $\dfrac{\partial^2 f}{\partial x_j x_i}(x)$ exists and is continuous for all $i, j = 1, ..., n$

1.32 Gradient

$\nabla f(x)$ is a column vector in $\mathbb{R}^n$ representing the gradient of f

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{bmatrix}$$

1.33 Directional Derivatives

$f'(x, r)$ is the directional derivative of x in direction v

$$\begin{aligned} f'(x, v) &= \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t} \\ &= \langle \nabla f(x), v \rangle \end{aligned}$$

1.34 Hessian

$\nabla^2 f(x)$ is a matrix in $\mathbb{R}^{n\times n}$ consisting of the second-order partial derivatives of f

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

$$\left[\nabla^2 f(x)\right]_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

- If f is C''-smooth, then $\partial^2 f(x)$ is symmetric

1.35 Directional Derivative Approximation Theorem

If f is C'-smooth, then $\lim\limits_{h\to 0} \dfrac{f(x+h) - f(x) - \langle \nabla f(x), h\rangle}{||h||} = 0$

- If f is C'-smooth, then the directional derivative of x in direction h represents the gradient of f in direction h
- Alternatively, we can write $f(x+h) - f(x) - \langle \nabla f(x), h\rangle = o(||h||)$
  - $f(x) = o(t)$ is notationally equivalent to $\lim\limits_{t\to 0} \dfrac{f(x)}{t} = 0$

1.36 Best Linear Approximation

If f is C'-smooth, then the best linear approximation of f centered at x is given by

$$g(y) = f(x) + \langle \nabla f(x), y - x\rangle$$

$$g(x+h) = f(x) + \langle \nabla f(x), h\rangle$$

1.37 Best Linear Approximation Error

The error equation for the best linear approximation of f centered at x is given by

$$\underbrace{f(x+h)}_{\text{function value}} = \underbrace{f(x) + \langle \nabla f(x), h\rangle}_{\text{linear approximation}} + \underbrace{o(||h||)}_{\text{error value}}$$

1.38 Mean Value Theorem

If f is C''-smooth, then for any $x, y \in \mathbb{R}^n$, there exists $z \in [x, y]$ such that

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\left\langle \nabla^2 f(z)(y-x), y - x\right\rangle$$

1.39 Taylor's Theorem

If f is C''-smooth, then

$$\underbrace{f(y)}_{\text{function value}} = \underbrace{f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\left\langle \nabla^2 f(x)(y-x), y - x\right\rangle}_{\text{quadratic approximation}} + \underbrace{o\left(||y-x||^2\right)}_{\text{error value}}$$

2 Section Two

2.1 Global Minimizers

Let $f : S \to \mathbb{R}$ where $S \subseteq \mathbb{R}^n$. Then $\bar{x} \in S$ is a global minimizer of f over S if $f(\bar{x}) \leq f(x)$ for all $x \in S$

- $\bar{x} \in S$ is a strict global minimizer of f over S if $f(\bar{x}) < f(x)$ for all $x \in S \backslash \{\bar{x}\}$
- $f(\bar{x})$ is the minimal value of f

2.2 Local Minimizers

Let $f : S \to \mathbb{R}$ where $S \subseteq \mathbb{R}^n$. Then $\bar{x} \in S$ is a local minimizer of f over S if there exists $r > 0$ such that $f(\bar{x}) \leq f(x)$ for all $x \in S \cap B(\bar{x}, r)$

- $\bar{x} \in S$ is a strict local minimizer of f over S if there exists $r > 0$ such that $f(\bar{x}) < f(x)$ for all $x \in S \cap B(\bar{x}, r)$

2.3 Critical Points

$\bar{x}$ is a critical point of a differentiable $f : \mathbb{R}^n \to \mathbb{R}$ if $\triangledown f(\bar{x}) = 0$

- A critical point can correspond to a local maximum, local minimum, or inflection point

2.4 Convex Functions

A C''-smooth function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\triangledown^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^n$

2.5 Coercive Functions

A function $f : \mathbb{R}^n \to \mathbb{R}$ is coercive if $\lim\limits_{i \to \infty} f(x_i) = +\infty$ for any $x_i \in \mathbb{R}^n$ where $||x_i|| \to +\infty$

- A quadratic function f is coercive if and only if $\triangledown^2 f(x)$ is positive definite

2.6 Principal Minors

If $A \in \mathbb{R}^{n \times n}$, then the determinant of the top-left $k \times k$ submatrix of A is the k^{th} principal minor, denoted as $\Delta_n(A)$

2.7 Recognizing Positive Definite and Semidefinite Matrices

- Let $\lambda_{\min}(A)$ be the minimal eigenvalue of a matrix A
  - $A \succ 0$ if and only if $\lambda_{\min}(A) > 0$
  - $A \succeq 0$ if and only if $\lambda_{\min}(A) \geq 0$
- $A \succ 0$ if and only if $\Delta_1(A), \Delta_2(A), ..., \Delta_n(A) > 0$
  - The test for positive semidefinite matrices requires that all principal minors of A be non-negative

2.8 First-Order Conditions

Let $\bar{x}$ be a local minimizer of $f : \mathbb{R}^n \to \mathbb{R}$. If f is differentiable at $\bar{x}$, then $\triangledown f(\bar{x}) = 0$

- Otherwise, $f(\bar{x} - t\triangledown f(\bar{x})) < f(\bar{x})$ for all small $t > 0$

2.9 Second-Order Conditions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be C''-smooth

- If $\bar{x}$ is a local minimizer of f, then $\triangledown f(\bar{x}) = 0$ and $\triangledown^2 f(\bar{x}) \succeq 0$
- If $\triangledown f(\bar{x}) = 0$ and $\triangledown^2 f(\bar{x}) \succ 0$, then $\bar{x}$ is a local minimizer of f

2.10 Sufficient Conditions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then the following are equivalent

- $\bar{x}$ is a local minimizer of f
- $\bar{x}$ is a global minimizer of f
- $\bar{x}$ is a critical point of f

2.11 Additional Theorems

- If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and coercive, then f attains its infimum
- If $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and coercive, and S is a closed set, then f attains its infimum over S
- $A \succeq 0$ if and only if there exists a lower triangular matrix L such that $A = LL^T$
  - L can be found via Cholesky factorization, which is beyond the scope of this class

2.12 Quadratic Functions

A quadratic function over $\mathbb{R}^n$ is a function of the form

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c$$
$$f(x_1, ..., x_n) = \frac{1}{2}\sum_{i,j} A_{ij}x_ix_j + \sum_i b_ix_i + c$$

where $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$

- $x^T Ax = x^T \left(\frac{A + A^T}{2}\right) x$
  - We can always assume A is symmetric since we can always express A as $\frac{A + A^T}{2}$
- The first-order derivative is given by $\triangledown f(x) = Ax + b$
- The second-order derivative is given by $\triangledown^2 f(x) = A$

2.13 Quadratic Functions Theorem

Let $f(x) = \frac{1}{2}x^T Ax + b^T x + c$ where $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Then

- x is critical if and only if $Ax + b = 0$
- f has a strict global minimizer if and only if $A \succ 0$
- f has a global minimizer if and only if $A \succeq 0$ and $b \in \mathrm{Range}(A)$
- If x satisfies $Ax + b = 0$, then x is a global minimizer
- f is coercive if and only if $A \succ 0$

3 Section Three

3.1 Least Squares

Given an inconsistent system of equations $Ax = b$, the least squares solution is an approximate solution that minimizes the squared norm of the residual $r = Ax - b$

$$\begin{aligned} f(x) &= \frac{1}{2}||Ax - b||_2{}^2 \\ &= \frac{1}{2}x^T(A^TA)x - (A^Tb)^Tx + \frac{1}{2}b^Tb \end{aligned}$$

$$\nabla f(x) = A^TAx - A^Tb$$

$$\nabla^2 f(x) = A^TA$$

- Least squares functions always have minimizers, represented by the solution of $\nabla f(x) = 0$
  - $A^TAx - A^Tb = 0$ always has a solution
  - $\nabla^2 f(x)$ is always positive semidefinite

3.2 Applications: Linear Fitting

Suppose we have data points $(s_i, t_i) \in \mathbb{R}^n \times \mathbb{R}$ for $i = 1, ..., m$. We want to find $x \in \mathbb{R}^n$ such that $t_i \approx s_i{}^Tx$ for all $i = 1, ..., m$. Then x represents the minimizer to the least squares function

$$f(x) = \frac{1}{2}||Sx - t||_2{}^2$$

where $S = \begin{bmatrix} s_1{}^T \\ \vdots \\ s_m{}^T \end{bmatrix}$ and $t = \begin{bmatrix} t_1 \\ \vdots \\ t_m \end{bmatrix}$

- Linear fitting finds the straight line of best fit through the dataset

3.3 Applications: Nonlinear Fitting

Suppose we have data points $(s_i, t_i) \in \mathbb{R} \times \mathbb{R}$ for $i = 1, ..., m$. We want to find a degree d polynomial $p(s_i) = a_0 + a_1s_1 + ... + a_ds_i{}^d$ such that $t_i \approx p(s_i)$. Then the coefficients $a = [a_0\ a_1\ ...\ a_m]$ represent the minimizer to the least squares function

$$f(a) = \frac{1}{2}||Sa - t||_2{}^2$$

where $S = \begin{bmatrix} 1 & s_1 & \dots & s_1{}^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & s_m & \dots & s_m{}^d \end{bmatrix}$ and $t = \begin{bmatrix} t_1 \\ \vdots \\ t_m \end{bmatrix}$

- Nonlinear fitting finds the polynomial line of best fit through the dataset

3.4 Applications: Regularized Least Squares

Nonlinear fitting tends to overfit data when given high enough degree d. To avoid this, we add a p function $R(x)$ to reinforce certain behaviors. This gives us the regularized least squares function

$$f(a) = \frac{1}{2}\underbrace{||Sa - t||_2{}^2}_{\text{fidelity}} + \lambda \underbrace{R(a)}_{\text{prior}}$$

where $S = \begin{bmatrix} 1 & s_1 & \dots & s_1{}^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & s_m & \dots & s_m{}^d \end{bmatrix}$ and $t = \begin{bmatrix} t_1 \\ \vdots \\ t_m \end{bmatrix}$ and $\lambda \geq 0$

- Typical choices for the prior function are
  - $R(a) = ||a||_2{}^2 = \sum_{j=0}^{m} a_j{}^2$
  - $R(a) = ||Da||_1 = \sum_{i=1}^{k} |(Dx)_i|$ for some $D \in \mathbb{R}^{k \times d}$
    * Forces many of $(Da)_i$ to be zero
  - $R(a) = \frac{1}{2}||Da||_2{}^2 = \frac{1}{2}\sum_{i=1}^{k} (Dx)_i{}^2$ for some $D \in \mathbb{R}^{k \times d}$
    * Forces all of $(Da)_i$ to be small
    * Minimizer is represented by the solution of $\left(S^T S + \lambda D^T D\right) a - A^T b = 0$

3.5 Applications: Denoising

Suppose we have data points $b_i = x_i + \omega_i$ for $i = 1, ..., m$ where x_i is the truth value and ω_i is some noise. We want to find the truth value x_i such that the line of best fit through x_i is a polynomial function. Then $x = [x_1 \; x_2 \; ... \; x_m]$ represents the minimizer to the least squares function

$$\begin{aligned} f(x) &= \frac{1}{2}||b - x||_2{}^2 + \frac{1}{2}\lambda||Lx||_2{}^2 \\ &= \frac{1}{2}\sum_{i=1}^{m} (b_i - x_i)^2 + \frac{1}{2}\lambda \sum_{i=1}^{m-1} (x_i - x_{i+1})^2 \end{aligned}$$

where $L = \begin{bmatrix} 1 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(m-1) \times m}$ and $b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$

- Minimizer is represented by the solution of $\left(I + \lambda L^T L\right) x - b = 0$
- The larger λ is, the smoother the denoised data becomes

3.6 Applications: Trend Filtering

Suppose we have data points $b_i = x_i + \omega_i$ for $i = 1, ..., m$ where x_i is the truth value and ω_i is some noise. We want to find the truth value x_i such that the line of best fit through x_i is

- A piecewise constant function. Then $x = [x_1\ x_2\ ...\ x_m]$ represents the minimizer to the least squares function

$$\begin{aligned} f(x) &= \frac{1}{2}||b - x||_2{}^2 + \frac{1}{2}\lambda||D^{(1)}x||_1 \\ &= \frac{1}{2}\sum_{i=1}^{m}(b_i - x_i)^2 + \frac{1}{2}\lambda\sum_{i=1}^{m-1}|x_i - x_{i+1}| \end{aligned}$$

where $D^{(1)} = \begin{bmatrix} 1 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & -1 \end{bmatrix} \in \mathbb{R}^{(m-1)\times m}$ and $b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$

- A piecewise linear function. Then $x = [x_1\ x_2\ ...\ x_m]$ represents the minimizer to the least squares function

$$\begin{aligned} f(x) &= \frac{1}{2}||b - x||_2{}^2 + \frac{1}{2}\lambda||D^{(2)}x||_1 \\ &= \frac{1}{2}\sum_{i=1}^{m}(b_i - x_i)^2 + \frac{1}{2}\lambda\sum_{i=1}^{m-2}|x_i - 2x_{i+1} + x_{i+2}| \end{aligned}$$

where $D^{(2)} = \begin{bmatrix} 1 & -2 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & -2 & 1 \end{bmatrix} \in \mathbb{R}^{(m-2)\times m}$ and $b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$

- A piecewise quadratic function. Then $x = [x_1\ x_2\ ...\ x_m]$ represents the minimizer to the least squares function

$$\begin{aligned} f(x) &= \frac{1}{2}||b - x||_2{}^2 + \frac{1}{2}\lambda||D^{(3)}x||_1 \\ &= \frac{1}{2}\sum_{i=1}^{m}(b_i - x_i)^2 + \frac{1}{2}\lambda\sum_{i=1}^{m-3}|x_i - 3x_{i+1} + 3x_{i+2} - x_{i+3}| \end{aligned}$$

where $D^{(3)} = \begin{bmatrix} 1 & -3 & 3 & -1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & -3 & 3 & -1 \end{bmatrix} \in \mathbb{R}^{(m-3)\times m}$ and $b = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$

4 Section Four

4.1 Line Search Method

Given an initial point $x \in \mathbb{R}^n$ in a C'-smooth function f, the line search method generates a sequence x_k for $k = 1, 2, ...$ such that $f(x_{k+1}) < f(x_k)$ and x_k approaches the local minimizer

$$x_{k+1} = x_k + t_k d_k$$

where $t_k \in \mathbb{R}$ is the step size and $d_k \in \mathbb{R}^n$ is the descent direction

4.2 Descent Direction

A non-zero vector d is a descent direction of a C'-smooth function f if the directional derivative of f along d is negative, that is $\langle f(x), d \rangle < 0$

- Given $\alpha \in (0, 1)$, there exists $\varepsilon > 0$ such that $f(x + td) < f(x) + \alpha t \langle f(x), d \rangle$ for all $t \in (0, \varepsilon)$
- Typical choices for the descent direction are
  - Gradient descent
    * $d_k = -\nabla f(x_k)$
  - Newton
    * $d_k = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$ where $\nabla^2 f(x_k) \succ 0$
  - Quasi-Newton
    * $d_k = H_k \nabla f(x_k)$ where H_k is an approximation of $-\left[\nabla^2 f(x_k)\right]^{-1}$

4.3 Step Size

A large step size allows x_k to approach the minimizer faster, while a small step size allows x_k to get closer to the minimizer

- Typical choices for the step size are
  - Constant
    * $t_k = \bar{t}$ for some $\bar{t} \in \mathbb{R}$
    * Decently fast, but not accurate ($x_k, x_{k+1}, ...$ might end up oscillating)
  - Exact line search
    * $t_k = \underset{t \geq 0}{\operatorname{argmin}} \ f(x_k + td_k)$
    * Finds the exact t value that minimizes $f(x_k + td_k)$
    * Fast and accurate, but calculating t_k for each iteration is impractical
  - Backtracking
    * Let $s > 0$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$
    * `set` $t \leftarrow s$
      `while` $f(x + td) \geq f(x) + \alpha t \langle \nabla f(x), d \rangle$
      $\quad t \leftarrow \beta t$
      `set` $t_k \leftarrow t$
    * Finds the largest approximate t value that minimizes $f(x_k + td_k)$
    * Decently fast and decently accurate

4.4 Condition Number

The condition number of a matrix A is defined as $\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$

- If $\kappa(A)$ is small, A is well-conditioned
- If $\kappa(A)$ is large, A is ill-conditioned

4.5 Gradient Descent

In gradient descent, the descent direction is opposite to the gradient such that $d_k = -\nabla f(x_k)$

$$x_{k+1} = x_k - \nabla f(x_k) t_k$$

- Gradient descent has linear convergence
  - Each iteration divides the error by a fixed constant
- The direction of motion is orthogonal to the contour line
  - $\langle x_{k+2} - x_{k+1},\ x_{k+1} - x_k \rangle = 0$ for all k
- If the contour lines of the graph are poorly scaled, then the direction of motion ends up zig-zagging excessively and the rate of convergence suffers

4.6 Lipschitz Property of the Gradient

Suppose f is C''-smooth. Then the following are equivalent

- $f \in C_L^{1,1}$ such that $||\nabla f(x) - \nabla f(y)||_2 \leq L||x - y||_2$ for all $x, y \in \mathbb{R}^n$
- $||\nabla^2 f(x)|| \leq L$ for all $x \in \mathbb{R}^n$
- $\max\limits_{i=1,\ldots,n} \left|\lambda_i\left(\nabla^2 f(x)\right)\right| \leq L$ for all $x \in \mathbb{R}^n$

4.7 Descent Lemma

Suppose $f \in C_L^{1,1}$. Then $f(y) \leq f(x) + \nabla f(x)^T(y - x) + \dfrac{L}{2}||x - y||^2$ for all $x, y \in \mathbb{R}^n$

- The descent lemma provides an upper bound for a quadratic function over the entire space

4.8 Sufficient Decrease Lemma

Suppose $f \in C_L^{1,1}$. Then $f(x) - f(x - t\nabla f(x)) \geq t\left(1 - \dfrac{Lt}{2}\right)||\nabla f(x)||^2$ for all $x \in \mathbb{R}^n$ and $t > 0$

4.9 Sufficient Decrease of the Gradient Descent

Suppose $f \in C_L^{1,1}$. Let $(x_k)_{k\geq 0}$ be the sequence generated by the gradient descent for solving

$$\min_{x\in\mathbb{R}^n} f(x)$$

with one of the following step size strategies

- Constant step size $t \in \left(0, \frac{2}{L}\right)$
- Exact line search
- Backtracking procedure with parameters $s \in \mathbb{R}_{++}$, $\alpha \in (0,1)$, and $\beta \in (0,1)$

Then $f(x_k) - f(x_{k+1}) \geq M||\nabla f(x_k)||^2$ where

$$M = \begin{cases} t\left(1 - \frac{tL}{2}\right) & \text{constant step size} \\ \frac{1}{2L} & \text{exact line search} \\ \alpha \min\left(s, \frac{2(1-\alpha)\beta}{L}\right) & \text{backtracking} \end{cases}$$

4.10 Convergence of the Gradient Descent

Suppose $f \in C_L^{1,1}$ and that there exists $m \in \mathbb{R}$ such that $f(x) > m$ for all $x \in \mathbb{R}^n$. Let $(x_k)_{k\geq 0}$ be the sequence generated by the gradient descent for solving

$$\min_{x\in\mathbb{R}^n} f(x)$$

with one of the following step size strategies

- Constant step size $t \in \left(0, \frac{2}{L}\right)$
- Exact line search
- Backtracking procedure with parameters $s \in \mathbb{R}_{++}$, $\alpha \in (0,1)$, and $\beta \in (0,1)$

Then we have the following

- The sequence $(f(x_k))_{k\geq 0}$ is monotone decreasing
- For any $k \geq 0$, $f(x_{k+1}) < f(x_k)$ unless $\nabla f(x_k) = 0$
- $\nabla f(x_k) \to 0$ as $k \to \infty$
- $\min\limits_{i=1,\ldots,k} ||\nabla f(x_k)||^2 \leq \frac{f(x_0) - \inf\limits_x f(x)}{M(k+1)}$

4.11 Upper Complexity Bound of Gradient Descent

To find a point x_i with $||\nabla f(x_i)|| \leq \varepsilon$, it suffices to perform $k = \frac{f(x_0) - \inf f}{M\varepsilon^2}$ iterations

4.12 Linear Rate Theorem

Suppose f is C''-smooth and $f \in C_L^{1,1}$ with $\lambda_{\min}\left(\nabla^2 f(x)\right) \geq \mu > 0$ for all $x \in \mathbb{R}^n$. Then gradient descent with constant step size $t_k = \frac{1}{L}$ satisfies

$$||x_{k+1} - \bar{x}||^2 \leq \left(1 - \frac{\mu}{L}\right) ||x_k - \bar{x}||^2$$

where $\bar{x}$ is a minimizer

- Then to find a point x_i with $||\nabla f(x_i)|| \leq \varepsilon$, it suffices to perform $k = \frac{L}{\mu} \ln\left(\frac{||x_0 - \bar{x}||^2}{\varepsilon}\right)$ iterations

4.13 Lower Complexity Bound of Gradient Descent

Suppose $f \in C_L^{1,1}$ with $\lambda_{\min}\left(\nabla^2 f(x)\right) \geq \mu > 0$ for all $x \in \mathbb{R}^n$. Then to find a point x_i with $||\nabla f(x_i)|| \leq \varepsilon$, it requires at least $k = \sqrt{\frac{L}{\mu}} \log\left(\frac{||x_0 - \bar{x}||^2}{\varepsilon}\right)$ iterations

5 Section Five

5.1 Newton's Method

In Newton's method, the descent direction is $d_k = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$ with step size $t = 1$

$$x_{k+1} = x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$$

- To calculate $\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$, we let $d_k = \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$ and solve the system of equations $\nabla^2 f(x_k) d_k = \nabla f(x_k)$

5.2 Disadvantages of Newton's Method

- Requires that the starting point is sufficiently close to the optimal point
- Requires us to calculate the Hessian $\nabla^2 f(x)$ at each iteration
- Requires us to solve a linear system of equations at each iteration

5.3 Upper Complexity Bound of Newton's Method

Suppose the starting point x_0 is sufficiently close to $\bar{x}$. To find a point x_k with $||x_k - \bar{x}|| \leq \varepsilon$, it suffices to perform $k = \log\left(\log\left(\frac{c}{\varepsilon}\right)\right)$ iterations, where c is some constant

5.4 Quadratic Local Convergence of Newton's Method

Suppose f is C'-smooth and let $\bar{x}$ satisfy $f(\bar{x}) = 0$. Suppose that there exists $\mu, \varepsilon, L > 0$ such that

- $\left|\left|[\nabla f(x)]^{-1}\right|\right| \leq \frac{1}{\mu}$ for all $x \in B(\bar{x}, \varepsilon)$
- $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$ for all $x, y \in B(\bar{x}, \varepsilon)$

Let (x_k) be the sequence generated by Newton's method and let $\bar{x}$ be the unique minimizer of f over $\mathbb{R}^n$. Then

$$||x_{k+1} - \bar{x}|| \leq \frac{L}{2\mu}||x_k - \bar{x}||^2$$

If $||x_k - \bar{x}|| \leq \frac{\mu}{L}$, then $||x_k - \bar{x}|| \leq \frac{2\mu}{L}\left(\frac{1}{2}\right)^{(2^k)}$

5.5 Affine Invariance of Newton's Method

Affine invariance means that surfaces are considered the same under affine/linear transformations. Therefore Newton's method performs the same with functions $f(x)$ and $f(Ax)$

6 Appendix

6.1 Common Expressions in $\mathbb{R}^{m\times n}$

- Given $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times m}$, $||Ax||_2{}^2 = (Ax)^T Ax = x^T A^T Ax$
- Given $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$, $x^T Ax = \displaystyle\sum_{i,j} A_{ij} x_i x_j$
- Given $x \in \mathbb{R}^n$, $x^T x = \displaystyle\sum_{i=1}^{n} x_i{}^2$
- Given $x \in \mathbb{R}^n$ and $A_{ij} = x_i x_j$, $A = xx^T$
- $\mathrm{Null}(A^T A) = \mathrm{Null}(A)$
- $\mathrm{Range}(A^T A) = \mathrm{Range}(A^T)$