Optimisation Cheat Sheet

Chapter 2: Optimality

a. **Fermats Theorem and Stationary Point:** For $f : U \to \mathbb{R}$ where $U \subseteq \mathbb{R}^n$ then optimal point $x^* \in U^\circ \implies (\frac{\partial f}{\partial x_i}$ exists $\implies \nabla f(x^*) = 0)$. A point in the interior where partial derivatives are defined and $\nabla f(x^*) = 0$ is called a stationary point.

b. **Second Order Classification:** If

$$x^* \text{ is a local min/max } \implies \nabla^2 f(x^*) \succeq / \preceq 0 \quad \text{and} \quad \nabla^2 f(x^*) \succeq / \preceq 0 \implies x^* \text{ is a local min/max or saddle}$$

but

$$\nabla^2 f(x^*) \succ / \prec 0 \implies x^* \text{ is a local strict min/max}$$

It is important to note that being a strict min/max does not mean the hessian is PD or ND.

c. **Saddle:** A saddle is a stationary point which is neither a local min or max. $\nabla^2 f(x^*)$ indefinite $\implies x^*$ is a saddle but not the other way around because it could still have a positive semi-definite.

d. **Coercive:** By weierstrass theorem, any coercive function admits a global minimum point where a coercive function $f(x)$ is such that $f(x) \to \infty$ as $||x|| \to \infty$.

e. **Global Optimality:** $f \in C^2 \implies (\nabla^2 f(x) \succeq 0 \; \forall \; x \in \mathbb{R}^n \implies (x^* \text{ local min} \implies x^* \text{ global min}))$

f. **Quadratic Functions:** A quadratic function is $f(x) = x^T Ax + 2b^T x + c$ where A is symmetric and $\nabla f = 2Ax + 2b, \nabla^2 f = 2A$.

   a. x^* is a global min $\iff A \succeq 0 \; \wedge x^* = A^{-1}b$

   b. Coercive $\iff A \succ 0$

   c. Positive $\iff \begin{pmatrix} A & b \\ b^T & c \end{pmatrix} \succeq 0$

Chapter 3: LS

a. **Regular Least Squares:** The solution to the problem

$$\min_{x \in \mathbb{R}^n} ||Sx - b||^2$$

which has a unique solution $x_{ls} = (S^T S)^{-1} S^T b$ when $S^T S$ is invertible which happens when it has full rank. When it is not full rank, then there are many minimisers.

b. **Regularised Least Squares:** To add some regularity to the solution we instead solve the problem

$$\min_{x \in \mathbb{R}^n} ||Sx - b||^2 + \lambda ||Dx||^2$$

which always has a() solution(s) but it is unique and given by $x_{rls} = (S^T S + \lambda D^T D)^{-1} S^T b$ when $(S^T S + \lambda D^T D)$ is invertible.

c. **Denoising:** Denoising is a specific kind of regularisation where we penalise large jumps between adjacent points.

$$\min_{x \in \mathbb{R}^n} ||x - b||^2 + \lambda \sum_{i=2}^{m} (x_i - x_{i-1})$$

which has solution $x_d = (I + \lambda L^T L)^{-1} b$.

d. **Circle Fitting LS:**

Chapter 4: Descent Methods

1. **Descent Direction:** For continuously differential $f : \mathbb{R}^n \to R$, d is a descent direction of f if the directional derivative in the direction of d is negative. The optimal descent direction is given by $-\nabla f(x)/||\nabla f(x)||$.

2. **Stepsize:** There are three choices in this course. Constant, exact and backtracking.

   a. **Constant stepsize:** $t^k = \bar{t}$ for all steps k.

   b. **Exact line search:** $t^k \in \arg\min_{t \geq 0} f(x^k - t \nabla f(x))$. The descent direction at each iteration in gradient descent with exact line search is orthogonal to the descent direction at the previous iteration.

   c. **Backtracking:** Parameters s, the initial step size, α which controls how much we want to decrease at each step and β, how much we scale the step size while we dont meet the following sufficient decrease condition

$$f(x^k) - f(x^{k+1}) \geq \alpha t^k ||\nabla f(x^k)||^2$$

3. **Gradient Descent:** For regular gradient descent we need to provide a tolerance parameter ϵ and an initial starting point x_0. Then we iteratively update the point in the direction of $-\nabla f(x^k)$ using a step size selection procedure above. Gradient descent converges for $C_L^{1,1}(\iff ||\nabla^2 f||$ bounded) functions which are bounded below.

4. **Scaled and Gauss-Newton:** Given a tolerence and intial point, the scalred gradient descent method just scales the descent direction by a matrix D^k at iteration k and sets $x^{k+1} \leftarrow x^k = -t^k D^k \nabla f(x^k)$

   The Gauss Newton method has the following scaling matrix

   $$D^k = \frac{1}{2}\left(J(x^k)^T J(x^k)\right)^{-1}$$

   but there are many alternatives for scaling, the main motivation is to make the problem better conditioned.

5. **Kaczmarz and SGD:** The Kaczmarz algorithm is a way to solve a linear system, $Ax = b$ and the solution x is attained by fitting each x_i to b_i. In the same spirit, for non-linear least squares, the SGD algorithm fits

Chapter 5: Convex Functions

1. **Convex Function:** A function over a convex set is convex if

   $$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \quad \text{for all } x, y \in C \text{ and } \alpha \in [0,1]$$

   and first a second order characterisations are the important gradient inequality

   $$f(x) + \nabla f(x)^T (y - x) \leq f(y) \quad \text{for all } x, y \in C$$

   and the hessian condition which is

   $$\nabla^2 f(x) \succeq 0 \quad \text{for all } x, y \in C.$$

   Obviously for all these conditions, we need to impose the appropriate differentiability requirements.

2. **Operations which preserve convexity:**

   a. Scalar multiplication and sums.

   b. Affine transformation of input: If f is convex over C then $g(y) = f(Ax+b)$ is convex over $D = \{x \in \mathbb{R}^n : Ax + b \in C\}$.

   c. Composition with non decreasing convex scalar function, ie $g : I \to \mathbb{R}$ such that $f(C) \subset I$ then $g \circ f$ is convex over C.

   d. Pointwise maximum of family of convex functions is convex.

   e. Partial minimisation of a convex function if $f : C \times D \to \mathbb{R}$ then $g(x) = \min_{y \in D} f(x, y)$ is convex.

3. **Simple Theorems about Convex functions**

   a. **Level Set Convex:** The level set of a function denoted $\text{Lev}(f, \alpha) = \{x \in C : f(x) = \alpha\}$ is convex.

   b. **Convex function is continuous:**

   c. **Existence of directional derivative of convex function:** The directional derivative $f'(x; d) = \lim_{t \downarrow 0}(f(x + td) - f(x))/t$ exists.

   d. **No maximum in interior, if convex and compact, then maximum is at the extreme points.**

Chapter 6: Convex Optimisation

1. **Basic Theorems:** Local optimal points are global optimal points and the set of optimal points is convex (if strict, then it is a singleton).

2. **Optimisation over a convex set:** For a C^1 function over a convex set we say it is stationary if the directional derivative from x^* is positive, ie. $\nabla f(x^*)^T (x - x^*) \geq 0$ for all $x \in C$. We trivially have that a local minimum is a stationary point. We want this definition because $\nabla f(x^*)$ is not necessarily 0 at the optimal point (at a boundary).

3. **Orthogonal Projection Operator:** This operator returns the closes point in a closed convex set to to the argument and is given by

   $$P_C(x) = \arg\min\{||y - x||^2 : y \in C\}$$

   and for any non empty closed convex subset of $\mathbb{R}^n$, this uniquely exists. Furthermore, an equivalent characterisation is

   $$z = P_C(x) \iff (x - z)^T (y - z) \leq 0 \text{ for all } y \in C$$

   and this basically says that the angle between $y - z$ and $x - z$ is more than $\pi/2$ for any $y \in C$. This is important because we can characterise stationary points as any point such that the orthogonal projection of any point in the direction of the gradient from that point is the point, or more concretely

   $$x^* = P_C(x^* - s\nabla f(x^*)$$

4. **Gradient Projection Method:** Basically gradient descent and but we project the outcome at each iteration to the convex set. It is more useful if the projection is analytic and we do not have to numerically solve it at each iteration. Because we know that a stationary point is a fixed point to the above equation, we use that as a stopping criterion, ie

$$G_L(x) = L\left(x - P_C\left(x - \frac{s}{L}\nabla f(x)\right)\right)$$

5. **Frank Wolfe:** When the orthogonal projection is unavailable, we can use the Frank Wolfe algorithm

Chapter 7: Optimality Conditions

1. **Technincal Results:**

   a. Separation Theorem: For a closed convex set C and $y \notin C$, there exists $p \in \mathbb{R}^n \setminus 0$ such that $p^T x \leq \alpha$ and $p^T y > \alpha$ for all $x \in C$. Basically Hahn Banach in finite dimensions.

   b. Farkas Lemma: For $c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$, $Ax \leq 0$ and $c^T x > 0$ exclusively or $A^T y = c$ and $y \geq 0$ where $\geq, \leq$ denotes component-wise inequalities.

   c. Farkas Lemma 2: $\left(Ax \leq 0 \implies c^T x \leq 0\right) \implies \exists\, y \in \mathbb{R}^m_+ : A^T y = c$

   d. Gordons Alternative: $A^T p = 0$ for $p \in \mathbb{R}^m_+ \setminus 0$ has a solution or $Ax < 0$ has a solution.

2. **Linear KKT Conditions:** The KKT conditions basically say that the gradient of the objective function is a linear combination of the gradients of the active constraints (the weights of all other constraints are 0).

   a. KKT conditions for Linearly Constrained Problems - Necessary Optimality Conditions: For continuously differentiable $f : \mathbb{R}^n \to \mathbb{R}$ and the problem, $\min f(x)$ for $x \in \{x : Ax \leq b\}$ we have

   $$\nabla f(x^*) = -\sum_{i=1}^{m} \lambda_i a_i \quad \text{and} \quad \lambda_i(a_i^T x^* - b_i) = 0$$

   where x^* is a local minimum, $\lambda_i \geq 0$ and a_i is the i-th row of A.

   b. KKT Conditions for Convex Linearly Constrained Problems - Necessary and Sufficient Optimality Conditions: When we add the convexity requirement, we get the leftward implication of the above statement. It is important to note that even if a C^1 function satisfies the KKT conditions above at a point, the point might not be a local minimum.

   c. KKT conditions for Linearly Constrained Problems: We can also add equality constraints to the continuously differentiable problem to get the problem, $\min f(x)$ for $x \in \{x : Ax \leq b \wedge Cx = d\}$ for $A \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times n}$. Particularly we have

   $$x^* \text{ is local min} \implies \exists \lambda \in \mathbb{R}^m_+ \text{ and } \mu \in \mathbb{R}^p : \nabla f(x^*) = -A^T\lambda - C^T\mu \quad \text{and} \quad \lambda_i(a_i^T x^* - b_i) = 0$$

   Furthermore, if we have convexity, we get the leftward implication aswell.

3. **Non-Linear KKT Conditions**:

   a. Feasible descent direction: Basically a direction you can go in which decreses the objective and is also contained in the convex set, C. So, for $x \in C$, $d \in \mathbb{R}^n : \nabla f(x)^T d < 0$ and $x + \epsilon d \in C$.

   b. For the continuously differentiable problem, $\min f(x)$ for $x \in \{x : g_i(x) \leq 0 \text{ for } i \in \{1, \ldots, m\}\}$ where $f, g_i \in C^1(C)$ we have that if x^* is a local minimum then there is no descent direction $d \in \mathbb{R}^n$ such that $\nabla f(x^*)^T d < 0$ and $\nabla g_i(x^*)^T d < 0$ for all active constraints. In effect there is no direction where we reduce all activate constrains and also reduce the objective.

   c. Sufficiency of the KKT conditions for convex optimization problems: For the convex C^1 problem

   $$\begin{aligned} \min \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0 \quad \text{for } i = 1, \ldots, m \\ & h_i(x) \leq 0 \quad \text{for } i = 1, \ldots, p \\ & s_i(x) = 0 \quad \text{for } i = 1, \ldots, q \end{aligned}$$

   with C^1 convex functions f, g and affine h, we have that if there exists $\lambda \in \mathbb{R}^m_+, \eta \in \mathbb{R}^p_+$ and $\mu \in \mathbb{R}$ such that

   $$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*) + \sum_{i=1}^{p} \eta_i \nabla h_i(x^*) + \sum_{i=1}^{q} \mu_i \nabla s_i(x^*) = 0 \quad \text{and} \quad \lambda_i g_i(x^*) = \eta_i h_i(x^*) = 0$$

   then x^* is an optimal point.

   d. Necessary KKT conditions under the generalized **Slater**'s condition: Furthermore, if x^* is an optimal point and there exists $\hat{x}$ such that

   $$\begin{aligned} g_i(\hat{x}) &< 0 \quad \text{for } i = 1, \ldots, m \\ h_i(\hat{x}) &\leq 0 \quad \text{for } i = 1, \ldots, p \\ s_i(\hat{x}) &= 0 \quad \text{for } i = 1, \ldots, q \end{aligned}$$

   then we get the rightward implication.

Chapter 8: Duality

1. **Langrangian and Dual Objective:** For the **primal** problem given by

$$\begin{aligned} & f^* = \min f(x) & \\ \text{subject to} \quad & g_i(x) \leq 0 & \text{for } i = 1, \dots, m \\ & h_i(x) = 0 & \text{for } i = 1, \dots, p \end{aligned}$$

   we have the lagrangian and the dual as

$$L(x, \lambda, \mu) = f(x) + \sum_{i=0}^{m} \lambda_i g_i(x) + \sum_{j=0}^{p} \mu_j h_j(x) \quad \text{and } q(\lambda, \mu) = \min_{x \in X} L(x, \lambda, \mu)$$

   where $\lambda \geq 0 \in \mathbb{R}^m$ and the **dual** problem is given by

$$\begin{aligned} & q^* = \max q(\lambda, \mu) \\ \text{s.t.} \quad & (\lambda, \mu) \in \text{dom}(q) := \{(\lambda, \mu) \in \mathbb{R}^m_+ \times \mathbb{R}^p : q(\lambda, \mu) > -\infty\} \end{aligned}$$

   Two import important facts about dual problem is that $\text{dom}(q)$ is convex and q is concave on it.

2. **Weak and Strong Duality:** For the primal and dual problem above we have that $q^* \leq f^*$. This might not be so useful in practice because q^* might no exist or it might be $-\infty$.

   Additionally, if we have only inequality constraints which are convex,

$$\begin{aligned} & f^* = \min f(x) & \\ \text{subject to} \quad & g_i(x) \leq 0 & \text{for } i = 1, \dots, m \\ & x \in X & \end{aligned}$$

   where f, g, X are **convex**, and there exists $\hat{x} \in X : g_i(\hat{x}) < 0$ (Slater's condition), then $|f^*| < \infty \implies q^*$ exists and $q^* = f^*$.

3. **Complementary Slackness:** The complementary slackness theorem is useful because it tells us that if the optimal values of the primal and dual are equal $f^* = q^*$, then we can check solutions are optimal if they satisfy

$$x^* = \arg\min_{x \in X} L(x, \lambda^*) \quad \text{and} \quad \lambda_i^* g_i(x) = 0$$

4. **General Strong Duality Conditions:** For

$$\begin{aligned} \min \quad & f(x) & \text{for convex } f \\ \text{subject to} \quad & g_i(x) \leq 0 & \text{for convex } g_i \ i = 1, \dots, m \\ & h_i(x) \leq 0 & \text{for affine } h_i \ i = 1, \dots, p \\ & s_i(x) = 0 & \text{for affine } s_i \ i = 1, \dots, q \\ & x \in \text{ convex } X & \end{aligned}$$

   then if there is $\hat{x} \in \text{int}(X)$ which satisfies the slates condition then $|f^*| < \infty \implies q^* = f^*$.

5. **Dual Algorithms:** Primal object f can be convex/non-convex but dual will always be concave.

   a. **Augmented Lagrangian:** The augmented Lagrangian for $\min_{x \in X}\{f(x) : Ax = b\}$ is given by

$$L_{\alpha_k}(x, \mu) = L(x, \mu) + \frac{\alpha_k}{2} ||Ax - b||^2$$

   and the augmented lagrangian method solves the problem for convex $f(x)$ by iteratively solving the augmented Lagrangian for the optimal $\mu \leftarrow \mu^k + \alpha_k(Ax - b)$.

   b. **ADMM:** For the problem $\min_{x,z \in X,Z}\{f(x) + g(z) : Ax + Bz = c\}$. Basically the above but you alternative between x and z between each update of μ.