CPSC 406 Review Notes

Reese Critchlow

Linear Algebra Review

Orthogonality: A $n \times m$ matrix is said to be orthonormal if its columns are pairwise orthonormal:

$$Q = [q_1 | \cdots | q_m].$$

It follows that then, for an orthonormal matrix, then:

$$Q^T Q = I_m.$$

In addition, if $n = m$, that is if Q is square, then Q is said to be orthogonal. From this, it follows that:

$$Q^{-1} = Q^T \qquad\qquad Q^T Q = QQ^T = I_n.$$

It is to be also noted that orthogonal transformations preserve lengths and angles.

Nonsingularity: A matrix is said to be **nonsingular** if it has a matrix inverse and is square. A square matrix is nonsingular iff its determinant is nonzero.

Condition Number: The condition number of an $n \times n$ positive definite matrix H is

$$\kappa(H) = \frac{\lambda_{\max}(H)}{\lambda_{\min}(H)} \geq 1.$$

It is said that a matrix is ill-conditioned if $\kappa(H) \gg 1$. If f is twice continuously differentiable, the condition number of f at solution x^* is given by:

$$\kappa(f) = \kappa(\nabla^2 f(x^*))$$

Linear Least Squares

A basic linear least squares problem has the form:

$$\min_x \|Ax - b\|_2^2.$$

In essence, linear least squares seeks to find the vector x that, when multiplied by the matrix A, returns the closest result to b. The ij-th entry of the A matrix can be interpreted as the i-th observation of the j-th independent variable.

Normal Equations: We can write the least squares problem as a function f of x, thus, the solution to the least squares problem must be a stationery point of f:

$$\begin{aligned} x^* = \arg\min_x f(x) := \frac{1}{2}\|Ax - b\|_2^2 \\ \implies \nabla f(x) = A^T - A^T b \\ \nabla f(x^*) = 0 \iff A^T A x^* - A^T b = 0 \iff A^T A x^* = A^T b \end{aligned}$$

If A is full rank, then the solution x^* is unique.

Geometric Interpretation: If the $n \times m$ matrix A has range range(A), and for some vector $b \in \mathbb{R}^m$ then the vector Ax^*, where x^* is the solution to the least squares problem is the projection of b onto the range(A). Hence, we can also define the residual r to be vector which is the difference between Ax^* and b, $r = b - Ax^*$. Hence, it must be that $r \in \text{null}(A^T)$, such that $A^T r = 0$.

QR Factorization

We can obtain the QR factorization for an $m \times n$ matrix, with $m > n$:

$$A = [Q_1|Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

Where:

- Q is orthogonal
- R_1 is [right] upper triangular
- $\text{range}(Q_1) = \text{range}(A)$
- $\text{range}(Q_2) = \text{range}(A)^{\perp} \equiv \text{null}(A^T)$.

We can use the QR factorization to solve $n \times n$ nonsingular matricies:

$$x = A^{-1}b = R^{-1}Q^T b$$

This can also be used to solve least squares problems. Due to the condition number, it is said that QR is a more numerically stable solution rather than the normal equations approach.

Singular Value Decomposition (SVD)

For any $m \times n$ matrix A with rank r:

$$A = U\Sigma V^T = [u_1|u_2|\cdots|u_r] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_r^T \end{bmatrix}$$

A simple interpretation of the components of the SVD are as follows:

- U is a basis for $\text{range}(A)$
- V is a basis for $\text{range}(A^T)$
- Σ contains all of the roots of the eigenvalues of A.

It is also nice to define two other norms for matricies given the SVD:

- Spectral Norm: $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \sigma_1$
- Frobenious Norm: $\|A\|_F = \sqrt{\text{trace}(A^T A)} = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$

We can say that the SVD decomposes any matrix A with rank r into a sum of rank-1 matricies. Hence, we can describe the best rank-k approximation by the following:

$$A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T.$$

We can also say that the full SVD provides orthogonal bases for all four fundamental subspaces for an $m \times n$ matrix:

- $\text{range}(A) = \text{span}\{u_1, \ldots, u_r\}$

- $\text{null}(A^T) = \text{span}\{u_{r+1}, \ldots, u_m\}$
- $\text{range}(A^T) = \text{span}\{v_1, \ldots, v_r\}$
- $\text{null}(A) = \text{span}\{v_{r+1}, \ldots, v_n\}$

Minimum norm least-squares solution: Building off of the prior result of the fundamental subspaces, we obtain that:

$$\overline{x} = Vy = \sum_{r}^{j=1} \frac{u_j^T b}{\sigma_j} v_j, \;\; \sigma_j y_j = \begin{cases} \overline{b}_j/\sigma_j & j = 1:r \\ 0 & j = r+1:n \end{cases}$$

Regularized Least Squares

Regularized Least Squares is motivated by multi-objective optimization problems, where one must choose some x to minimize $f_1(x)$ and $f_2(x)$, but they do not get small together. Commonly, the solution space is divided into two parts, one containing possible solutions, and one containing impossible solutions. The boundary between these two sets is called the Pareto Frontier.

Weighted-Sum Objective: Commonly, the approach to a multi-objective optimization is to weight the sum of objectives:

$$\min_x \alpha_1 f_1(x) + \alpha_2 f_2(x)$$

Hence, the negative ratio of the two αs ends up becoming the slope of the Pareto Frontier at each given solution point on the curve $-\left(\frac{\alpha_1}{\alpha_2}\right)$.

Tikhonov Regularization/Ridge Regression: This form of the least squares problem is generally employed for the case that the standard least-squares problem is ill-posed, and thus requires some sort of bias. It can also be applied for a case in which one would like to have a solution with particular characteristics be favoured. It is as follows:

$$\min_x \frac{1}{2}\|Ax - b\|^2 + \frac{1}{2}\lambda\|Dx\|^2$$

Where:

- $\|Dx^2\|$ is the regularization penalty (often $D = I$)
- λ is the positive regularization parameter

Hence, we can say that an equivalent expression for the objective is:

$$\|Ax - b\|^2 + \lambda\|Dx\|^2 = \left\| \begin{bmatrix} A \\ \sqrt{\lambda}D \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2.$$

Hence, the normal equations then become:

$$(A^T A + \lambda D^T D)x = A^T b.$$

Gradients, Linearizations, and Optimality

For some function $f : \mathbb{R}^n \to \mathbb{R}$, then $x^* \in \mathbb{R}^n$ is a:

- **Global Minimizer** if $f(x^*) \leq f(x),\ \forall x.$
- **Strict Global Minimizer** if $f(x^*) \leq f(x),\ \forall x.$
- **Local Minimizer** if $f(x^*) \leq f(x),\ \forall x \in \epsilon\mathbf{B}(x^*).$
- **Strict Local Minimizer** if $f(X^*) < f(x),\ \forall x \in \epsilon\mathbf{B}(x^*).$

1-Dimensional Optimization: Standard Calc 1 definitions:

- **Local Minimizer**: $f'(x) = 0 \wedge f''(x) > 0$
- **Local Maximizer**: $f'(x) = 0 \wedge f''(x) < 0$

Multidimensional Optimization

Directional Derivatives: A directional derivative of $f : \mathbb{R}^n \to \mathbb{R}$ at $x \in \mathbb{R}^n$ in the direction $d \in \mathbb{R}^n$ is:

$$f'(x; d) = \lim_{\alpha \to 0^+} \frac{f(x + \alpha d) - f(x)}{\alpha}$$

Descent Directions: A nonzero vector d is a descent direction of f at x if:

$$f(x + \alpha d) < f(x),\ \forall \alpha \in (0, \epsilon) \text{ for some } \epsilon > 0 \iff f'(x; d) < 0$$

Gradients: If $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, the gradient of f at x is the vector:

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

It is also implied that: $f'(x;d) = \nabla f(x)^T d$. And, if $\|d\| = 1$, then the projection of the gradient onto d is given by $f'(x;d) \cdot d$.

Level Set (definition): The α-level set of f is the set of points of x where the function value is at most α. A direction d points into the level set $[f \leq f(x)]$ if $f'(x;d) < 0$.

Multidimensional Conditions

Matrix Definiteness: Let A be an $n \times n$ matrix with $A = A^T$ (symmetric).

- **Positive Semidefinite**: A is positive semidefinite ($H \succeq 0$) if:
  - $x^T Ax \geq 0 \ \forall \ x \in \mathbb{R}^n$
  - For a diagonal matrix $D \succeq 0 \iff d_i \geq 0 \ \forall i$
  - All eigenvalues are greater than or equal to zero.
  - $A = R^T R$ for some $n \times n$ matrix R.
- **Positive Definite** ($A \succ 0$) if
  - $x^T Ax > 0 \ \forall \ 0 \neq x \in \mathbb{R}^n$
  - For a diagonal matrix $D \succ 0 \iff d_i > 0 \ \forall i$
  - All eigenvalues are strictly greater than zero.
  - $A = R^T R$ for some nonsingular $n \times n$ matrix R.
- **Indefinite** if:
  - $\exists x \neq y \in \mathbb{R}^n$ such that $x^T A(x) > 0 \wedge y^T Ay < 0$.

Hessians: For $f : \mathbb{R}^n \to \mathbb{R}$, twice continuously differentiable, the **Hessian** of f at $x \in \mathbb{R}^n$ is the $n \times n$ symmetric matrix:

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

Quadratic Functions: Quadratic functions take the following forms:

- $f(x) = \frac{1}{2} x^T Hx + b^T x + \gamma$
- $\nabla f(x) = Hx + b$
- $\nabla^2 f(x) = H$

Directional Second Derivatives: The directional second derivative of f at x in the direction d is given by:

$$f''(x;d) = \lim_{\alpha \to 0^+} \frac{f'(x + \alpha d; d) - f'(x;d)}{\alpha} = d^T \nabla^2 f(x) d$$

Linear and Quadratic Approximations:

- Linear Approximation: $f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x)$
- Quadratic Approximation: $f(x + d) = f(x) + \nabla f(x)^T d + \frac{1}{2} d^T \nabla^2 f(x) d + o(\|d\|^2)$

Sufficient Conditions for Optimality: For $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable and $\overline{x} \in \mathbb{R}^n$ stationery ($\nabla f(\overline{x}) = 0$), if:

- $\nabla^2 f(\overline{x}) \succeq 0 \implies \overline{x}$ is a local min.
- $\nabla^2 f(\overline{x}) \preceq 0 \implies \overline{x}$ is a local max.
- $\nabla^2 f(\overline{x}) \succ 0 \implies \overline{x}$ is a *strict* local min.
- $\nabla^2 f(\overline{x}) \prec 0 \implies \overline{x}$ is a *strict* local max.
- $\nabla^2 f(\overline{x})$ indefinite, then $\overline{x}$ is undetermined. (test does not tell us anything)

Nonlinear Least Squares

We define the nonlinear least-square problem (NLLS) to be:

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{2}\|r(x)\|_2^2 \qquad\qquad r : \mathbb{R}^n \to \mathbb{R}^m$$

Where r is the residual function is given by:

$$r(x) = \begin{bmatrix} r_1(x) \\ r_2(x) \\ \vdots \\ r_m(x) \end{bmatrix}, \qquad J(x) = \begin{bmatrix} \nabla r_1(x)^T \\ \nabla r_2(x)^T \\ \vdots \\ \nabla r_m(x)^T \end{bmatrix}, \qquad \nabla f(x) = J(x)^T r(x)$$

This reduces to linear least-squares when r is affine.

Descent Methods

Gradient Descent: Intialization: Choose $x_0 \in \mathbb{R}^n$ and tolerance $\epsilon > 0$.
Iterations:

1. Choose step size $\alpha^{(k)}$ to approximatley minimize $\phi(\alpha) = f(x^k - \alpha \nabla f(x^k))$
2. Update: $x^{(k+1)} = x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)})$
3. Stop: if $\|\nabla f(x^{(k)})\| < \epsilon$.

Scaled Descent: Scaled descent is motivated by the zig-zagging behaviour of gradient descent, where exact linesearch implies that descent "steps" must be orthogonal to each other. Specifically, if the condition number κ is large, $\kappa \gg 1$, then the zig-zagging is exacerbated. Hence, scaled gradient seeks to make a change of variables $x = Sy$ for some nonsingular S. Hence, for an original minimization problem:

$$\min_x f(x) \qquad\qquad f : \mathbb{R}^n \to \mathbb{R}$$

The change of variables implies:

$$\min_y g(y) = f(Sy).$$

The gradient of g is given by:

$$\nabla g(y) = S^T \nabla f(Sy).$$

And thus, we get that the scaled gradient method is:

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)} \qquad\qquad d^{(k)} = -SS^T \nabla f(x^{(k)}).$$

Where $SS^T \succ 0$ (given earlier in Matrix Deefiniteness).

Hence, we can give a method for scaled gradient descent:

1. Choose some scaling matrix $D^{(k)} = SS^T \succ 0$.
   - Remark: choosing a scaling matrix S is generally done in a way which makes the condition number of the rescaled matrix as close to 1 as possible ($\kappa(\nabla^2 g) \approx 1$).
   - Some common scalings include:

$$S^{(k)}(S^{(k)})^T = \begin{cases} (\nabla f(x^{(k)}))^{-1} & \text{Newton } (\kappa = 1) \\ (\nabla f(x^{(k)}) + \lambda I)^{-1} & \text{Damped Newton} \\ \mathbf{Diag}(\frac{\partial f(x^{(k)})}{\partial x_i^2})^{-1} & \text{diagonal scaling} \end{cases}$$

2. Compute $d^{(k)} = -D^{(k)} \nabla f(x^{(k)})$.
3. Choose stepsize $\alpha^{(k)} > 0$ through linesearch.
4. Update $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$.

Gauss Newton for NLLS:
Objective: $\min_x f(x) = \frac{1}{2}\|r(x)\|_2^2$
Step: $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$
Procedure:

1. Compute residual $r_k = r(x^{(k)})$ and Jacobian $J_k = J(x^{(k)})$
2. Compute step $d^{(k)} = \text{argmin}_d \|J_k d + r_k\|^2$, ($d^{(k)} = -J_k \backslash r_k$).
3. Choose stepsize $\alpha^{(k)} \in (0, 1]$ via linesearch on $f(x)$.
4. Update $x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)}$.
5. Stop if $\left\|r(x^{(k+1)})\right\| < \epsilon$ or $\left\|\nabla f(x^{(k)})\right\| = \left\|J_k^T r_k\right\| < \epsilon$.

Step Size Selection:

- **Exact** (hard/expensive except for quadratic case): $\alpha^{(k)} \in \text{argmin}_{\alpha \geq 0} \phi(\alpha)$.
  - Quadratic Case: Choose $\alpha^* = -\dfrac{\nabla f(x)^T d}{d^T H d}$
- **Constant** (cheap/easy, but requires analysis of f): $\alpha^{(k)} = \overline{\alpha} > 0, \ \forall k$.
  - Quadratic Case: Choose $\alpha \in \left(0, \dfrac{2}{\lambda_{\max}(H)}\right)$
- **Approximate** [Backtracking/Armijo] (cheap, no analysis):
  1. Set $\alpha^{(k)} > 0$
  2. Until $f(x^{(k)} + \alpha^{(k)} d^{(k)} < f(x^k) + \mu\alpha f'(x; d))$, $\mu \in (0, 1)$.
     - $\alpha^{(k)} \leftarrow \alpha^k \cdot \rho$, $\rho \in (0, 1)$.
  3. Return $\alpha^{(k)}$

Lipschitz Smooth Functions: A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be L-Lipschitz smooth if:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \ \ \forall x, y \in \mathbb{R}^n$$

For quadratic functions, we can say that $L = \lambda_{\max}(H)$.
Second-order L-smooth characterization: If f is twice continuously differentiable, then f is L-Lipschitz smooth iff its Hessian is bounded by L $\forall x \in \mathbb{R}^n$:

$$LI - \nabla^2 f(x) \succeq 0.$$

Cholesky Factorization

Cholesky factorization is another way of obtaining a LU-like decomposition, however, it only works if the matrix is positive definite. In relation to the LU decomposition, it uses $(1/3)n^3$ flops vs $(2/3)n^3$ for LU factorization. It is as follows:

$$A = R^T R$$

Where R is some positive definite upper triangular matrix.

Newton's Method

Newton's method arises as a product from the second order approximation of f at $x^{(k)}$. Hence, we get two forms of Newton's method:

Pure Newton's Method:

$$x^{(k+1)} = x^{(k)} + d_N^{(k)} \qquad\qquad H_k d_N^{(k)} = -\nabla f(x^{(k)})$$

Damped Newton's Method:

$$x^{(k+1)} = x^{(k)} + \alpha d_N^{(k)} \qquad\qquad \alpha \in (0, 1]$$

For Newton's method to converge, we require that $\nabla^2 f(x^{(k)}) \succ 0 \; \forall k$ to ensure descent. However, it is important to note that this does not always hold, such as in the case that $\lambda_{\min}(H_k)$ is small.

We can also use the Cholesky factorization for Newton's method. This process looks like:

1. Choose $\tau = 0$
2. Find the Cholesky factorization: $(H_k + \tau I) = R^T R$
    - If the Cholesky fails, increase τ and repeat.
3. Solve $R^T R d_N^k = -g_k$.

Linear Constraints

We can define a linearly constrained problem to be one such that:

$$\min_{x \in \mathbb{R}^n} \{f(x) \,|\, Ax = b\}$$

Where:

- $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function
- A is $m \times n$, $m < n$
- $b \in \mathbb{R}^m$
- A has full rank.

Feasible Set: A feasible set is the set of all vectors which satisfy the equation $Ax = b$:

$$\mathcal{F} = \{x \in \mathbb{R}^n \,|\, Ax = b\}.$$

We can also represent the feasible set in an alternative fashion:

$$\mathcal{F} = \{x \in \mathbb{R}^n \,|\, Ax = b\} = \left\{\overline{x} + Zp \,|\, p \in \mathbb{R}^{n-m}\right\}$$

Where:

- $\overline{x}$ is a particular solution ($A\overline{x} = b$)
- Z is a basis for the null space of A ($AZ = 0$)

Hence the problem becomes unconstrained in $n - m$ variables:

$$\min_{p \in \mathbb{R}^{n-m}} f(\overline{x} + Zp).$$

We can then apply any optimization to obtain the solution p^*, and $x^* = \overline{x} + Zp^*$ is the solution to the original problem.