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.

Second Half of the Course

Convex Sets

Definition: A set $C \subseteq \mathbb{R}^n$ is said to be *convex* if for any $\mathbf{x}, \mathbf{y} \in C,\ \lambda \in [0,1]$, the point $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in C$.

Convex Hull: The convex hull of a set of points $\mathcal{S}$ contains all convex combinations of points in $\mathcal{S}$:

$$\text{conv}(\mathcal{S}) = \left\{ \sum_{i=1}^{k} \theta_i x_i \,|\, x_i \in \mathcal{S},\, \sum_{i=1}^{k} \theta_i = 1, \theta_i \geq 0 \right\}$$

Equivalently, we can say that $\mathcal{C} \subset \mathbb{R}^n$ is convex if it contains all convex combinations of its elements, i.e. $\mathcal{C} = \text{conv}(\mathcal{C})$.

Subspace: $\mathcal{S} \subset \mathbb{R}^n$ is a subspace if it contains all linear combinations of points in the set, i.e.

$$\alpha x + \beta y \in \mathcal{S}, \forall x, y \in \mathcal{S}, \forall \alpha, \beta \in \mathbb{R}$$

For any $m \times n$ matrix A, its range and nullspace are subspaces of $\mathbb{R}^n$:

$$\textbf{range}(A) = \{Ax \,|\, x \in \mathbb{R}^n\} \qquad \text{and} \qquad \textbf{null}(A^T) = \left\{z \,|\, A^T z = 0\right\}$$

Affine Sets: $\mathcal{L}$ is an affine set if it's a translated subspace, that is, for fixed $x_0 \in \mathbb{R}^n$ and subspace $\mathcal{S}$:

$$\mathcal{L} = \{x_0 + v \,|\, v \in \mathcal{S}\} \equiv x_0 + \mathcal{S}$$

Halfspaces and Hyperplanes: For some fixed nonzero vector $a \in \mathbb{R}^n$ and scalar β, a hyperplane is given by:

$$\mathcal{H} = \left\{x \,|\, a^T x = \beta\right\}$$

and a halfspace is given by:

$$\mathcal{H}_- = \left\{x \,|\, a^T x \leq \beta\right\}$$

- a is normal to the hyperplane.
- Hyperplanes are both **affine** and **convex**.
- Halfspaces are convex but **not affine**.

Cones: A set $\mathcal{K} \subset \mathbb{R}^n$ is a **cone** if $x \in \mathcal{K} \iff ax \in \mathcal{K},\ \forall \alpha \geq 0$.

Convex Cones: A convex cone is a cone that is also convex:

$$x, y \in \mathcal{K} \wedge \alpha, \beta \geq 0 \implies \alpha x + \beta y \in \mathcal{K}$$

Some examples of convex cones include:

- Nonnegative Orthant: $\mathbb{R}^n_+ = \{x \,|\, x_i \geq 0, i = 1, \ldots, n\}$
- Second-Order Cone: $\mathcal{L}^n_+ = \left\{ \begin{bmatrix} x \\ t \end{bmatrix} \in \mathbb{R}^{n+1} \,\middle|\, \|x\|_2 \leq t, x \in \mathbb{R}^n, t \in \mathbb{R} \right\}$
- Positive Semidefinite Cone

Operations that Preserve Convexity: Let $\mathcal{C}_1, \mathcal{C}_2$ be convex sets in $\mathbb{R}^n$.

- Nonnegative Scaling: $\theta\mathcal{C}_1 = \{\theta x \,|\, x \in C_1\},\ \theta \geq 0$.
- Intersection: $\mathcal{C}_1 \cap \mathcal{C}_2$.
- Sum: $\mathcal{C}_1 + \mathcal{C}_2 = \{x + y \,|\, x \in \mathcal{C}_1,\, y \in \mathcal{C}_2\}$.

Convex Polytopes: $\mathcal{S}$ is a **convex polytope** if it is the intersection of a fininte number of halfspaces:

$$\mathcal{S} = \bigcap_{i=1}^{m} \left\{ x \,|\, a_i^T x \leq \beta_i \right\} = \{ x \,|\, Ax \leq b \}$$

where:

$$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} \in \mathbb{R}^{m \times n} \qquad \text{and} \qquad b \in \mathbb{R}^m$$

n-dimensional simplex: Is the intersection of n halfspaces and a hyperplane:

$$\mathcal{C} = \left\{ x \,\middle|\, \sum_{i=1}^{n} = 1,\ x_i \geq 0 \right\}.$$

Convex Functions

Convexity: A function $f : C \to \mathbb{R}$ is **convex** if $\mathcal{C} \subset \mathbb{R}^n$ is convex for all $x, y \in \mathcal{C}$ and $\theta \in [0, 1]$:

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta) f(y).$$

Strict Convexity: A function f is strictly convex if the inequality is strict for $x \neq y$ and $\theta \in (0, 1)$:

$$f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta) f(y).$$

Concavity: f is concave if $(-f)$ is convex.

Examples of Convex Functions:

- Exponential: e^{ax}
- Powers: x^α over $x \geq$ for any $\alpha \geq 1$ or $\alpha \leq 0$
- Absolute Value: $|x|^\alpha$ for any $\alpha \geq 1$
- Norms: $\|x\|_p$ for any $p \geq 1$

Examples of Concave Functions:

- Powers: x^α over $x \leq 0$ for any $0 < \alpha < 1$
- Logarithm: $\log x$ over $x > 0$

Convex and Concave:

- Affine: $a^T x + \beta$ for any $a \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$

Restriction to Lines: $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if and only if

$$\phi(a) = f(x + \alpha d)$$

is convex over $\alpha \in \mathbb{R}$ for all points x and directions d.

Operations that Preserve Convexity:

- Nonnegative Scaling
- Sums
- Composition with Affine Function

Convex Optimization:

$$\min_{x \in \mathcal{C}} f(x), \ \ \mathcal{C} \subset \mathbb{R}^n \text{ convex}, \ \ f : \mathcal{C} \to \mathbb{R} \text{ convex}$$

If x^* is a **local minimizer**, it is also a **global minimizer**.

Convexity and Level Sets: Given the level set of $f : \mathbb{R}^n \to \mathbb{R}$ at level $\alpha \in \mathbb{R}$:

$$[f \leq \alpha] := \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}$$

If f is convex, then all level sets are convex.

First-Order Characterizations: Let $f : \mathcal{C} \to \mathbb{R}$ be differentiable over $\mathcal{C} \subset \mathbb{R}^n$. Then f is convex if and only if:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \forall x, y \in \mathcal{C}.$$

Second-Order Characterization: Let $f : \mathcal{C} \to \mathbb{R}$ be twice differentiable over $\mathcal{C} \subset \mathbb{R}^n$. Then f is convex if and only if:

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathcal{C}.$$

Projected Gradient Descent

For a set $C \subset \mathbb{R}^n$ closed convex, the **projection** of a point $x \in \mathbb{R}^n$ onto C is the point:

$$\mathbf{proj}_C(x) = \underset{z \in C}{\operatorname{argmin}} \|x - z\| .$$

It is also important to note that:

$$z = \mathbf{proj}_C(x) \iff z \in C \wedge (x - z)^T (y - z) \leq 0 \ \forall y \in C$$
$$z = \mathbf{proj}_C(x) \iff -\nabla g(z) = x - z \in \mathcal{N}_C(z)$$

Projection onto Affine Set: Given some set $C = \{z \in \mathbb{R}^n \mid Az = b\}$ for $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$, then:

$$\mathbf{proj}_C(x) = \operatorname{argmin} \left\{ \frac{1}{2} \|z - x\|^2 \mid Az = b \right\}.$$

Because $\mathcal{N}_C = \mathbf{range}(A^T)$, the optimality condition is such that:

$$x - \mathbf{proj}_C(x) \in \mathbf{range}(A^T)$$

Projected Gradient Descent: Projected gradient descent proceeds as follows:

$$\min_x \{f(x) \mid x \in C\},$$

for some $f : \mathbb{R}^n \to \mathbb{R}$ convex and smooth and $C \subset \mathbb{R}^n$.
Hence, the algorithm for this is as follows:

- Start from $x_0 \in C$
- For $k = 0, 1, 2, \ldots$
  - $g_k = \nabla f(x_k)$
  - Linesearch on $\phi(\alpha) = f(\mathbf{proj}_C(x_k - \alpha g_k))$
  - $x_{k+1} = \mathbf{proj}_C(x_k - \alpha_k g_k)$
  - Stop if $||x_{k+1} - x_k||$ is small.

Stationarity:

$$x^* \underset{x \in C}{\operatorname{argmin}} f(x) \iff x^* = \mathbf{proj}_C(x^* - \alpha \nabla f(x^*)) \ \forall \alpha > 0$$

By the projection theorem:

$$(x^* - \alpha \nabla f(x^*) - x^*)^T (z - x^*) \leq 0 \ \forall z \in C$$
$$-\alpha \nabla f(x^*)^T (z - x^*) \leq 0 \ \forall z \in C$$

Given the definition of a normal cone, we obtain that:

$$-\nabla f(x^*) \in \mathcal{N}_C(x^*).$$

Convex Optimality

Given some convex optimization problem:

$$\min_x \{f(x) \,|\, x \in C\},$$

we say that:

- if $C = \mathbb{R}^n$ the problem is **unconstrained**, so $\nabla f(x^*) = 0$.
- however, if $C \neq \mathbb{R}^n$, then this does not imply that $\nabla f(x^*) = 0$.

Normal Cone: The normal cone to the set $C \subset \mathbb{R}^n$ at the point $x \in C$ is the set:

$$\mathcal{N}_C(x) = \left\{d \in \mathbb{R}^n \,|\, d^T(z - x) \leq 0,\ \forall z \in C\right\}$$

Necessary and Sufficient Optimality: A point $x^* \in \text{argmin}_{x \in C} f(x)$ if and only if:

$$\nabla f(x^*)^T(x - x^*) \geq 0 \forall x \in C.$$

Using the definition of the normal cone we can deduce the equivalent condition:

$$-\nabla f(x^*) \in \mathcal{N}_C(x^*).$$

Interior Point: It is implied that points in the interior have empty sets as normal cones.
This implies for unconstrained optimality that $\nabla f(x^*) = 0$.

Normal Cone to an Affine Set: Given some

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

Then for any $x \in C$, define the translated set as:

$$C_x = \{z - x \,|\, z \in C\}.$$

Then,

$$\mathcal{N}_C(x) = \textbf{range}(A^T) = A^T y$$

So x^* is optimal iff $Ax^* = b$ and $-\nabla f(x^*) = A^T y$ for some $y \in \mathbb{R}^n$.

And such, the vector $y \in \mathbb{R}^m$ contains all of the **Lagrange Multipliers** for each constraint: $a_i^T x = b_i$.

Convergence of Gradient Descent

- Quadratic functions with $f(x) = \frac{1}{2}x^T Ax + b^T x + \gamma$, with $A \succeq 0$, has $L = \|A\|_2 = \lambda_{\max}(A)$
- If f twice continuously differentiable, then f is L-smooth if and only if, for all x:

$$\nabla^2 f(x) \preceq LI \iff \|\nabla^2 f(x)\|_2 \leq L.$$

Descent Lemma: If f is L-smooth, then for all x, z, then f is **globally majorized** by a quadratic approximation.

$$f(z) \leq f(x) + \nabla f(x)^T(z - x) + \frac{L}{2}\|z - x\|^2$$

From prior it is known that we get decreasing objecitive values if we choose some stepsize $\alpha \in \left(0, \frac{2}{L}\right)$.

Strong Convexity

A function $f : \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex (with $\mu > 0$) if for all x, y:

$$f(z) \geq f(x) + \nabla f(x)^T (z - x) + \frac{\mu}{2} \|z - x\|^2$$

If f is **twice continuously differentiable**, then f is μ-strongly convex if and only if for all x:

$$d^T \nabla^2 f(x) d \geq \mu \|d\|^2 \ \ \forall d \in \mathbb{R}^n \iff \nabla^2 f(x) \succeq I\mu$$

Alternatively, one can say that a function f is μ-strongly convex if and only if for all x,

$$g(x) = f(x) - \frac{\mu}{2} \|x\|^2$$

This also implies that Tikhonov regularization induces strong convexity.
Distance to Solution: Some important lemmas arise out of this.

1. If f is L-smooth, then for all x and all minimizers x^* with $f^* = f(x^*)$:

$$\frac{1}{2L} \|\nabla f(x)\|^2 \leq f^* \leq \frac{L}{2} \|x - x^*\|^2$$

2. If f is μ-strongly convex, then for all x and all minimizers x^* with $f^* = f(x^*)$:

$$\frac{\mu}{2} \|x - x^*\|^2 \leq f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$$

The sum of all of these conclusions implies that the Hessian eigenvalues can be bounded from above and below:

$$\mu I \preceq \nabla^2 f(x) \preceq LI$$

We can also deduce the number of iterations T it would take to get down to a desireable error:

$$T \leq \frac{L}{\mu} \log \left(\frac{f_0 - f^*}{\epsilon} \right)$$

Stochastic Gradient Descent

For some least-squares gradient descent problem, where one would have that:

$$\nabla f(x) = \sum_{i=1}^{N} \nabla f_i(x)$$

However, we could take samples $j = 1, \ldots, N$ that occur with equal probability $\frac{1}{N}$ to get that:

$$f(x) = \frac{1}{N} \sum_{j=1}^{N} f_j(x) = \mathbb{E}_j f_j(x)$$

Hence, tha gradient is that:

$$\nabla f(x) = \frac{1}{N} \sum_{j=1}^{N} \nabla f_j(x) = \mathbb{E}_j \nabla f_j(x)$$

Hence, we randomly sample a small patch of observations $\mathcal{B} \subseteq \{1, \ldots, N\}$, then:

$$g_{\mathcal{B}}(x) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathbb{B}} \nabla f_i(x) \implies \mathbb{E}_{\mathcal{B}} g_{\mathcal{B}}(x) = \nabla f(x)$$

This is called a "stochastic approximation".

Stochastic Gradient Descent:

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

So each step we have that:

$$x_{k+1} = x_k - \alpha_k \nabla g_k$$

Where:

$$g_k := \frac{1}{B_k}\sum_{i\in B_k} \nabla f_i(x_k)$$

Where B_k is a batch of uniformly random iid samples from $\{1,\ldots,N\}$

- Step length α_k often called "Learning Rate" in this context.
- Need to assume mean-squared error in stochastic approximation is bounded:

$$\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] = \mathbb{E}\left[\|g_k\|^2\right] - \|\nabla f(x)\|^2 \leq \sigma^2$$

Convergence in Expectation (constant steplength): It works out such that:

$$\mathbb{E}f_{k+1} \leq \mathbb{E}f_k - \frac{\alpha}{2}\mathbb{E}\|\nabla f_k\|^2 + \frac{\alpha^2\sigma^2 L}{2}$$

If we sum over $k = 0, 1, 2, \ldots, T$ and recurse, rearranging, we get that:

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f^*)}{\alpha T} + \frac{\alpha\sigma L}{2}$$

Linear Programming

Conventional Program: Given $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n$:

$$\begin{aligned} \underset{x}{\text{minimize}}\ & c^T x \\ \text{subject to}\ & Ax = b \\ & x \geq 0 \end{aligned}$$

Diet Problem:

- Minimum-cost diet
- x_i represents how many servings of food group i to eat
- c_i gives cost of 1 serving of food from group i
- $a_i^T x = b_i$ encodes nutritional recommendations
- $x \geq 0$ since you can't eat negative food

Geometry:
Consider a linear program to be in a polyhedron:

$$\mathcal{P} = \{x \mid Ax \leq b\}$$

Then, an extreme point $x \in \mathcal{P}$ is if there does not exist two vectors $y, z \in \mathcal{P}$ such that:

$$x = \lambda y + (1-\lambda)z \text{ for any } \lambda \in (0,1).$$

Verticies: $x \in \mathcal{P}$ is a vertex of $\mathcal{P}$ if there exists a vector $c \neq 0$ such that:

$$c^T x < c^T y \text{ for all } y \in \mathcal{P}, y \neq x$$

- Given a vertex x, find c such that $c^T x < c^T y$ for all $y \in \mathcal{P}$, $y \neq x$.
- Given a vector c, find x such that $c^T x < c^T y$ for all $y \in \mathcal{P}, y \neq x$:

$$\underset{x}{\text{minimize}} \ c^T x \text{ subject to } x \in \mathcal{P}.$$

Active Constraints: Define $\mathcal{B}$ as the set of **active** or **binding** constraints (at x^*):

- Active Constraints: $a_i^T x^* = b_i, i \in \mathcal{B}$
- Inactive Feasible Constraints: $a_i^T x^* < b_i, i \in \mathcal{N}$
- Inactive Infeasible Constraints $a_i^T x^* > b_i, i \notin \mathcal{B} \cup \mathcal{N}$

Hence, the subset of active constraints are as follows:

$$A_{\mathcal{B}} = \overline{A} = \begin{bmatrix} a_{i_1}^T \\ \vdots \\ a_{i_k}^T \end{bmatrix} \qquad b_{\mathcal{B}} = \overline{b} = \begin{bmatrix} b_{i_1} \\ \vdots \\ b_{i_k} \end{bmatrix} \qquad \mathcal{B} = \{i_1, \ldots, i_k\}$$

Basic Solutions: x^* is a **basic solution** if one of the following equivalent conditions hold:

- $a_{i_1}, a_{i_2}, \ldots, a_{i_n}$ are linearly independent
- $\overline{A} x^* = \overline{b}$ has a unique solution
- $\text{rank}(\overline{A}) = n$

Basic Feasible Solution: x^* is a basic solution and $x^* \in \mathcal{P}$.

The following are equivalent:

- x^* is a vertex
- x^* is an extreme point
- x^* is a basic feasible solution

Unbounded Directions: $\mathcal{P}$ contains a **half-line** if there exists $d \neq 0$, x_0 such that:

$$x_0 + \alpha d \in \mathcal{P} \text{ for all } \alpha \geq 0$$

- $\mathcal{P}$ unbounded $\iff$ $\mathcal{P}$ contains a half-line
- $\mathcal{P}$ has no extreme points $\iff$ $\mathcal{P}$ contains a line
- $p^* = -\infty$ if and only if there exists a feasible half line
- $p^* = +\infty$ if and only if $\mathcal{P} = \emptyset$
- p^* is finite if and only if $X^* \neq \emptyset$

Simplex Method

Assume standard form of an LP problem:

$$\begin{aligned} &\underset{x}{\text{minimize}} \ c^T x \\ &\text{subject to } \ Ax = b, x \geq 0 \end{aligned}$$

Assuming that:

- A has full row rank (no redundant rows)
- The LP is feasible
- All basic feasible solutions are nondegenerate

We also define two variable index sets:

- $\mathcal{B} = \{\beta_1, \beta_2, \ldots, \beta_m\}$
- $\mathcal{N} = \{\eta_1, \eta_2, \ldots, \eta_{n-m}\}$

Feasible Directions: A direction d is **feasible** at $x \in \mathcal{P}$ if there exists $\alpha > 0$ such that:

$$x + \alpha d \in \mathcal{P}$$

Constructing Feasible Directions: Given some $x \in \mathcal{P}$ and $Ax = b$, $x \geq 0$, require that for all $\alpha \geq 0$ that:

$$b = A(x + \alpha d) = Ax + \alpha Ad = b + \alpha Ad$$

Thus, we require that $Ad = 0$. Suppose that x is a basic feasible solution, such that:

$$0 = Ad = \begin{bmatrix} B & N \end{bmatrix} \begin{bmatrix} d_b \\ d_n \end{bmatrix} = Bd_B + Nd_N \implies Bd_B = -Nd_N$$

We then construct search directions by moving a **single** nonbasic variable $\eta_k \in \mathcal{N}$:

$$d_N = e_k \text{ and } Bd_B = -Ne_k = -a_{\eta_k}$$

This goes on with swapping variables inside and outside of the basic set.

Choosing Swaps:

- For choosing $d_N = e_p$, choose p such that $z_{n_p} < 0$ (most negative) (for some $z = c - A^T y$, $B^T y = c_B$)
- Some basic variable β_q must exit the basis, choose: $q = \underset{q | d_{\beta_q} < 0}{\operatorname{argmin}} \left(-\frac{x_{\beta_q}}{d_{\beta_q}} \right)$

Hence, the new basic and nonbasic variables are as follows:

- $\mathcal{B} \leftarrow \mathcal{B} \setminus \{\beta_q\} \cup \{\eta_p\}$
- $\mathcal{N} \leftarrow \mathcal{N} \setminus \{\eta_p\} \cup \{\beta_q\}$

Converting Linear Programs to Standard Form

A generic polyhedron has the form:

$$\mathcal{P} = \left\{ x \;\middle|\; \begin{matrix} Ax = b \\ Cx \leq d \end{matrix} \right\}$$

However, it follows that a standard-form polyhedron has the form:

$$\mathcal{P} = \left\{ x \;\middle|\; \begin{matrix} Ax = b \\ x \geq 0 \end{matrix} \right\} \text{ with } b \geq 0.$$

Converting to Standard Form:

1. Positive b and positive d:
   (a) For $b_i < 0$, replace:
   $$a_i x = b_i \to (-a_i)x = (-b_i)$$
   (b) For $d_i < 0$, replace:
   $$c_i^T x \leq d_i \to (-c_i)^T x \geq (-d_i)$$
   $$c_i^T x \geq d_i \to (-c_i)^T x \leq (-d_i)$$

2. Free Variables

   - x_i is called a **free variable** if it has no constraints
   - There are no free variables in standard form – every variable must be nonnegative

   Converting Free Variables: Every free variable x_i is replaced with two new variables x_i' and x_i'':

   $$x_i = x_i' - x_i'',\ x_i' \geq 0 \text{ and } x_i'' \geq 0$$

   - x_i' encodes the positive part of x_i
   - x_i'' encodes the negative part of x_i
   - optimal solution must have that $x_i' \cdot x_i'' = 0$

3. Slack and Surplus

   For every inequality constraint of the form:

   $$c_i^T x \leq d_i \qquad\qquad (c_i^T x \geq d_i)$$

   Introduce a new **slack** (or **surplus**) variable s_i, replacing the inequality with two constraints:

   $$\begin{cases} c_i^T x + s_i = d_i \\ s_i \geq 0 \end{cases} \qquad\qquad \left(\begin{cases} c_i^T x - s_i = d_i \\ s_i \geq 0 \end{cases} \right)$$

In standard form, there are:

- n variables $(x_1, \ldots, x_n)$
- $m + n$ total constraints
  - m equality constraints $(Ax = b)$
  - n inequality constraints $(x \geq 0)$

For any basic solution x:

- The basic set $\mathcal{B}$ must have n elements
- Thus, exactly n of the constraints need to be active at x
- m equality constraints are always satisfied
- $n - m$ of the inequality constraints $x \geq 0$ should be "active"

Duality

Primal Problem:

$$\underset{x}{\text{minimize}}\ c^T x \text{ subject to } Ax = b,\ x \geq 0$$

- n variables, m constraints
- optimal solution x^*
- optimal value $p^* \equiv c^T x^*$

Relaxed Problem:

$$\underset{x}{\text{minimize}}\ c^T x + y^T(b - Ax) \text{ subj to } x \geq 0$$

- the relaxed problem provides a lower bound for p^*:

$$g(y) := \min_{x \geq 0} \left\{ c^T x + y^T (b - Ax) \right\} \leq c^T x^* + y^T (b - Ax^*) = c^T x^* = p^*$$

Tightest Lower Bound: Find y that solves

$$\underset{y}{\text{maximize}} \;\; g(y).$$

It follows that:

$$g(y) = \begin{cases} b^T y & \text{if } c - A^T y \geq 0 \\ -\infty & \text{otherwise} \end{cases}$$

Because we want to **maximize** $g(y)$, we must have that:

$$\begin{array}{rcl} \underset{y}{\text{maximize}} \;\; b^T y & \iff & \underset{y,z}{\text{maximize}} \;\; b^T y \\ \text{subject to} \;\; c - A^T y \geq 0 & & \text{subject to} \;\; A^T y + z = c, z \geq 0 \end{array}$$

This is the **dual** LP.

Weak Duality: Suppose that x is primal feasible:

$$Ax = b, \;\; x \geq 0$$

Suppose that (y, z) is dual feasible:

$$A^T y + z = c, \;\; z \geq 0$$

Then the primal objective is bounded below by the dual objective.

$$c^T x \geq y^T b.$$

Weak Duality Theorem: If (x, y, z) is primal/dual feasible, then:

- The primal value is an upper bound for the dual value
- The dual value is a lower bound for the primal value

Complementarity:

- If $x^T z = 0$ we say the bound is "tight".

Strong Duality Theorem: If an LP has an optimal solution, so does its dual, and the optimal values are equal, ie, $p^* = d^*$.

Theorem: The primal-dual triple (x, y, z) is optimal if and only if:

(1) $Ax = b, x \geq 0$

(2) $A^T y + z = c, z \geq 0$

(3) $x^T z = 0$

Matrix Games

Player X	Player Y
$\underset{x,\lambda}{\text{maximize}} \;\; \lambda$	$\underset{y,\nu}{\text{minimize}} \;\; \nu$
$\text{subject to} \;\; \lambda e \leq Ax$	$\text{subject to} \;\; \nu e \geq A^T y$
$e^T x, \; x \geq 0$	$e^T y, \; y \geq 0$

Multiplicative Weights Update Method

Given some game, where on round t:

- Player picks weights p_t in the **simplex**

$$p_t \in \Delta_n = \left\{ p \in [0,1]^n : \sum_{i=1}^{n} p_i = 1 \right\}$$

- Adversary picks losses $\ell_t \in [-1,1]^n$ for the experts
- By the end of the round, player loses $p_t(i) \cdot \ell_t(i)$ for stock/expert i:

$$\text{Loss on round } t = \sum_{i=1}^{n} \ell_t(i) \cdot p_t(i) = l_t^T p_t$$

We can evaluate the player through regret:

$$\text{Regret}(T) = \sum_{t=1}^{T} \ell_t^T p_t - \min_{i \in [n]} \sum_{t=1}^{T} \ell_t(i)$$

Note: Average regret goes to 0.