- Linear Regression: $\min_x f(x) := \sum_{i=1}^{n}(x^\top a_i - b_i)^2 = \|Ax-b\|^2$ for (a_i, b_i)
- SVM: $\min_x f(x) := \frac{1}{2}\|x\|^2$ s.t. $b_i\langle a_i, x\rangle \geq 1,\ \forall i$
- Cauchy-Schwarz: $-\|x\|\|y\| \leq x^\top y \leq \|x\|\|y\|$
- General Formulation: $\min_x f(x)$ s.t. $g_i(x) = 0, h_j(x) \leq 0, \quad \forall i \in \{1, \cdots, m\},\ j \in \{1, \cdots, p\}$
- Thm: If $\nabla f(x) = 0$, f is convex $\implies$ x is local (and global) optimal solution.
- Thm (Necessary): If x is a local optimal solution, it must be a stationary point.
- Thm: Every global optimal must be a local optimal solution.
- Thm (Sufficient): If x is stationary, $H_f(x)$ is PD/ND $\implies$ local min/max, is ID $\implies$ saddle, is PSD/NSD $\implies$ needs more investigation.
- Thm (Weistrass extreme-value): A continuous function on non-empty compact set must have global min and max.
- Thm: Coercive Function (if f is continuous, $\lim_{\|x\|\to\infty} f(x) = +\infty$) must have global min on non-empty S.

Convex Sets (conditions for $D \subseteq \mathbb{R}^n$:

- $x, y \in D \implies \lambda x + (1-\lambda)y \in D$
- Sub-level set: $D = \{x \in \mathbb{R}^n \mid f(x) \leq a\}$, f is convex
- Examples: Half Space ($H = \{x \in \mathbb{R}^n \mid a^\top x \leq b\}$), Closed Ball ($B(a,b) = \{x \in \mathbb{R}^n \mid \|x-a\| \leq b\}$), Polyhedral Set ($P = \{x \in \mathbb{R}^n \mid Ax \leq b\}$)
- Intersection of Convex Sets: Also Convex

Convex Functions:

- $\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ (Function below lining joining 2 points)
- Generally: $f\left(\sum_{i=1}^{n}\lambda_i x^{(i)}\right) \leq \sum_{i=1}^{n}\lambda_i f(x^{(i)})$
- $f(x) = \sup_{a,b\in\omega_f}(\phi_{a,b}(x) := a + b^\top x)$
- If f is C^1, $f(x) + \nabla f(x)^\top(y-x) \leq f(y)$ (Tangent below function)
- If f is C^1, $\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq 0, \forall x, y \in S$ (monotone gradient)
- If f is C^2, H_f is PSD
- Examples: $a^\top x + b, \|x\|, x^{2n}, \max\{x_1, \cdots, x_n\}, \exp(x), -\log x$
- : Linear combination with non-negative coefficients: Also convex

Positive Definite Matrices:

- $x^\top Ax \geq 0, \forall x$
- $\forall \lambda_i \geq 0$
- $\forall \Delta_k > 0$ (Test for PD)

Sets:

- Bounded: $\exists M$ s.t. $\|x\| \leq M$
- Closed: If complement is an open set.
- Compact: Closed + Bounded

Jacobi Coordinate Descent (start with x^0):

- $x_i^{k+1} = \arg\min_{x_i} f(w_{-i}^k)$ s.t. $w_{-i}^k = [x_1^k; \cdots; x_{i-1}^k; x_i^k; \cdots, x_n^k]$
- Slower than Gauss Siedel, but supports parallelization.

Gauss Siedel Coordinate Descent (start with x^0):

- $x_i^{k+1} = \arg\min_{x_i} f(w_{-i}^k)$ s.t. $w_{-i}^k = [x_1^{k+1}; \cdots; x_{i-1}^{k+1}; x_i^k; \cdots, x_n^k]$
- Faster than Jacobi, but doesn't support parallelization.

Coordinate Gradient Descent: $x_i^{k+1} = x_i^k - t_i^k \partial_{x_i} f(w_{-i}^k)$

Stochastic Gradient Descent:

- Target Loss Function $f(x) = \mathbb{E}_z F(x,z) = \mathbb{E}_{a,b} L(g(x,a), b)$
- Empirical Loss Function $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} F(x, z_i) = \frac{1}{n}\sum_{i=1}^{n} L(g(x, a_i), b_i)$
- GD: $x^{k+1} = x^k - t_k \frac{1}{n}\sum_{i=1}^{n}\nabla F(x, z_i)$; Each Step: $O(n)$
- SGD: $x^{k+1} = x^k - t_k \nabla F(x, z_J)$, j is randomly chosen; Each Step: $O(1)$
- Avg. SGD: Take $\frac{1}{T}\sum_{k=1}^{T} x^k$ in the later iterations
- $O(1/n)$ convergence

Choosing t_k:

- Conditions: $\sum_{k=1}^{\infty} t_k = \infty, \sum_{k=1}^{\infty} t_k^2 < \infty$
- Eg. $t_k = t_0 k^{-\alpha}, \alpha \in (0.5, 1]$

- Convex Problem: If f and S are convex, i.e., f, h_j are all convex C^1 and $g_i(x) = a_i^\top x - b_i$
- Def: $h_j \leq 0$ is active inequality constraint at x: $h_j(x) = 0$
- All equality constraints: active
- Def: $J(x) = \{j \in \{1, \cdots, p\} \mid h_j(x) = 0\}$
- Def: Regular Point (LICQ): If $x \in S$ (feasible) **and** the set $A_x = \{g_i(x) \forall i\} \cup \{h_j(x) : j \in J(x)\}$ is linearly independent i.e., $\text{rank}(A_x) = m + |J(x)|$
- Thm (Necessary): If x is local min for constrained optimization problem with C^1 f, g_i, h_j , x must be a 1st order KKT point. If C^2 f, g_i, h_j , x must be a 2st order KKT point.
- Thm (Sufficient): If x satisfies 2nd order KKT point condition, it must be a local min of f on S.
- Thm: If the problem is convex, $x \in S$ is KKT (1st order), x is global min of f on S.

KKT First Order Conditions for x:

- x is regular
- $\exists \lambda_1, \cdots, \lambda_m \in \mathbb{R}; \mu_1, \cdots, \mu_p \geq 0$ s.t. $\nabla f(x) + \sum_{i=1}^{m}\lambda_i \nabla g_i(x) + \sum_{j=1}^{p}\mu_j \nabla h_j(x) = 0; \mu_j = 0\ \forall j \notin J(x)$

Complementary Slackness Condition: $\mu_j h_j(x) = 0, \forall j$

KKT Second Order Conditions for x:

- x is KKT first order
- $H_L(x) = H_f(x) + \sum_{i=1}^{m}\lambda_i H_{g_i}(x) + \sum_{j=1}^{p}\mu_j H_{h_j}(x)$ is PSD on tangent space $T(x) \perp N(x)$
- Tangent Space: $T(x) = \{y \in \mathbb{R}^n \mid \nabla g_i(x)^\top y = 0, \nabla h_j(x)^\top y = 0; \forall i, \forall j \in J(x)\}$
- Normal Space: $N(x) = \text{span}(\nabla g_i(x), \nabla h_j(x); \forall i, \forall j \in J(x))$

If $H_L(x)$ is PD on $T(x)$, x is strict local min.

- Primal Problem (P): Original Constrained optimization problem.
- Lagrangian Function: $L(x, \mu, \lambda) = f(x) + + \sum_{i=1}^{m}\lambda_i g_i(x) + \sum_{j=1}^{p}\mu_j h_j(x)$
- Lagrangian Dual Function: $\theta(\lambda, \mu) = \min_{x\in X} L(x, \lambda, \mu)$
- Lagrangian Dual Problem (D): $\max_{\mu\geq 0}\theta(\lambda, \mu)$
- Duality Gap: $\Delta_g = \min_{x\in S} f(x) - \max_{\mu\geq 0}\theta(\lambda, \mu) \geq 0$

Important Theorems:

- $\theta(\lambda, \mu)$ must be a concave function if finite for all $(\lambda, \mu \geq 0)$ (but not necessarily C^1.
- Weal Duality Theorem: For $x \in S$, $\mu \geq 0$: $f(x) \geq \theta(\lambda, \mu)$. If $f(x') = \theta(\lambda', \mu')$, x' is optimal solution to (P) and (λ', μ') is optimal solution to (D)
- Strong Duality Theorem: If X is convex, (P) is convex, S is non-empty and $0 \in g(X) = \{g(x) : x \in X\} \implies \Delta_g = 0$

Linear Optimization:

- Problem: $\min f(x)$ s.t.$x^\top a_j - b_j \leq 0, \forall i$ or $Ax - b \leq 0$
- Optimal solution must be an extremal point, a vertex of the polygon formed by the constraints i.e., $x* \in V = \cup\{h_{j_1} = 0, \cdots, h_{j_d} = 0\}$
- Equivalent to solving $\max -\mu^\top b$ s.t. $\mu \geq 0, c^\top + \mu^\top A = 0$ by method of lagrangian duals.

Projected Gradient Descent

- Input: $x_0, \nabla f, t_k, S, Tol$
- Output: x_k s.t. $\|\nabla f(x)\| \leq$ Tol, $x_k \in S$
- Process (while $\|x_k - x_{k-1}\| \geq$ Tol):
  1. $v_k = -\nabla f(x_k)$
  2. $t_k = \arg\min_t f(x_k + tv_k)$ [Using Line Search]
  3. $x_{k+1} = \Pi_S(x_k + t_k v_k)$
  4. k = k + 1
- $\Pi_S(y) = \arg\min_x \frac{1}{2}\|x-y\|^2$ s.t. $x \in S$, solve using KKT. Alternatively, $x' = \Pi_S(y) \iff \langle y - x', x - x'\rangle \leq 0, \forall x \in S$
- Thm: If S is closed convex, $\Pi_S(y)$ must be unique for every y.
- Eg. 1: $S = \{\|x\| \leq a\} \implies \Pi_S(y) = y$ if $y \in S$, else $\dfrac{ay}{\|y\|}$
- Eg. 2: $S = \{a^\top x \leq b\} \implies \Pi_S(y) = y$ if $y \in S$, else $y - \dfrac{a^\top y + b}{\|a\|^2} a$

Penalty Method with Inequality Constraints:

- $S^- = \{x \in \mathbb{R}^n : h_j(x) < 0, \forall j\}$
- $P(x;\mu) = f(x) + \mu B(x)$, $B(x) = \sum_{j=1}^{p} \phi(-h_j(x)) = -\sum_{j=1}^{p} \log(-h_j(x))$
- Thm: Let x' is optimal solution s.t. $S^- \cap N\phi$ for any neighbourhood N around x'. Let x_μ be optimal solution for $\inf\{P(x;\mu) : x \in S^-, \mu > 0\}$. Limit of $x_m u$ must be the optimal solution of the original problem.
- Summary: $x_\mu = \arg\min_x P(x;\mu)$, $x' = \arg\min_{x\in S} f(x) = \lim_{\mu\to 0} x_\mu$
- Stopping Criterion: $p\mu_k \leq \epsilon$
- Algorithm (Logarithmic Barrier Method):
  - Input: $x_0, \mu_0, \nabla f, \nabla h, \rho, h, Tol$
  - Output: x_k s.t. $p\mu_k \leq$ Tol, $h(x_k) < 0$
  - Process (while $p\mu_k \geq$ Tol):
    1. $x = x_k$
    2. Inner While Loop (while $\|\nabla P(x,\mu_k)\| = \left\|\nabla f(x) - \mu_k \sum_{j=1}^{p} \dfrac{\nabla h_j(x)}{h_j(x)}\right\| >$ Tol:
       (a) $v = -\nabla f(x) + \mu_k \sum_{j=1}^{p} \dfrac{\nabla h_j(x)}{h_j(x)}$
       (b) $t = \arg\min_t f(x + tv)$ [Using Line Search]
       (c) $x = x + tv$
    3. $x_{k+1} = x; \mu_{k+1} = \rho\mu_k; k = k+1$

Penalty Method with Equality Constraints:

- $Q(x;\mu) = f(x) + \dfrac{1}{2\mu} B(x)$, $B(x) = \sum_{i=1}^{m} g_i^2(x)$
- Thm: Let x_μ be exact global minimizer of $Q(x;\mu)$. Limit of $x_m u$ must be the optimal solution of the original problem.
- Summary: $x_\mu = \arg\min_x Q(x;\mu)$, $x' = \arg\min_{x\in S} f(x) = \lim_{\mu\to 0} x_\mu$
- Stopping Criterion: $\|g(x)\| <\leq \epsilon$
- Algorithm (Quadratic Penalty Method):
  - Input: $x_0, \mu_0, \nabla f, \nabla g, \rho, g, Tol$
  - Output: x_k s.t. $\|\nabla g(x_k)\|, \|\nabla f(x_k)\| \leq$ Tol
  - Process (while $\|\nabla g(x_k)\| >$ Tol):
    1. $x = x_k$
    2. Inner While Loop (while $\|\nabla Q(x,\mu_k)\| = \left\|\nabla f(x) + \dfrac{1}{\mu_k}\sum_{i=1}^{m} g_i(x)\nabla g_i(x)\right\| >$ Tol:
       (a) $v = -\nabla f(x) - \dfrac{1}{\mu_k}\sum_{i=1}^{m} g_i(x)\nabla g_i(x)$
       (b) $t = \arg\min_t f(x + tv)$ [Using Line Search]
       (c) $x = x + tv$
    3. $x_{k+1} = x; \mu_{k+1} = \rho\mu_k; k = k+1$

Regularization in ML:

- $\min \mathbb{E}F(x,z) \to \min \mathbb{E}F(x,z) + \lambda R(x) \approx \min f(x) := \dfrac{1}{n}\sum_{i=1}^{n} F(x, z_i) + \lambda R(x) \iff \min \mathbb{E}F(x,z)$ s.t. $R(x) \leq R_0$
- $R(x)$ is high $\implies$ more model complexity. Purpose: To avoid overfitting + get unique solutions + improve identifiability.
- $R(x) = \|x\|_2^2$: l2/Tikhonov/Ridge; Smooth Convex [Linear Regression Solution: $x^* = (A^\top A + \lambda I)^{-1} A^\top b$]
- $R(x) = \|x\|_1 = \sum_{i=1}^{d} |x_i|$: l1/Lasso; Continuous, convex, non-differentiable; more sparsity
- $R(x) = \|x\|_0 = \#\{i : x_i \neq 0\}$: cardinality; Non-continuous, non-convex

Sub Gradient Method

- Def: v is subgradient of a convex function f at x if $f(y) \geq f(x) + \langle v, y - x\rangle, \forall y \in \mathbb{R}^n$
- Def: Subdifferential of f at x ($\partial f(x)$) is set of all subgradients of f at x. $\partial f(x) = \cap \Phi_y = \cap\{v \mid \psi_y(v) \leq 0\}; \psi_y(v) = f(x) + \langle v, y - x\rangle - f(y)$
- Thm: If f is continuous and convex, $x^* = \arg\min_x f(x) \implies 0 \in \partial f(x^*)$.
- Thm: $\partial f(x)$ is a convex set.
- Def: Convex hull: $C = \text{Conv}(S)$ if C is smallest convex set containing S. If $S = \{v_1, \cdots, v_k\} \implies C = \text{Conv}(S) = \left\{v = \sum_{i=1}^{k} \lambda_i v_i \mid \lambda_i \geq 0 \forall i; \sum_{i=1}^{k} \lambda_i = 1\right\}$

Strategies for calculating subdifferential:

- If f is differentiable at x, $\partial f(x) = \{\nabla f(x)\}$
- If $f(x) = \max\{f_1(x), \cdots, f_m(x)\}$, where f_i are all **convex** C^1. If $f_1(x') = \cdots = f_m(x')$, $\partial f(x') = \text{Conv}(\{\nabla f_1(x'), \cdots, \nabla f_m(x')\})$. [Therefore, $\partial f(a) = [-\rho, \rho], f = \rho|x - a|$]
- If f is differentiable at x, $\partial(f+g)(x) = \{\nabla f(x) + v \mid v \in \partial g(x)\}$
- $\{u + v \mid u \in \partial f(x), v \in \partial g(x)\} \subseteq \partial(f+g)(x)$
- If $\partial f(x) = \text{Conv}(\{u_1, \cdots, u_m\}), \partial g(x) = \text{Conv}(\{v_1, \cdots, v_n\})$, $\partial(f+g)(x) = \{\nabla u_i + v_j \mid i = 1, \cdots, m; j = 1, \cdots, n\}$.

Subgradient Descent:

- Input: $x_0, \partial f, t_k$
- Output: x_k s.t. $0 \in \partial f(x_k)$
- Process (while $0 \notin \partial f(x_k)$):
  1. Pick $v_k \in -\partial f(x_k)$
  2. $t_k = \arg\min_t f(x_k + tv_k)$ [Using Line Search]
  3. $x_{k+1} = \Pi_S(x_k + t_k v_k)$
  4. k = k + 1

Additional Notes:

- $\mathbb{E}(a^\top x - b)^2 = (x - x^*)^\top \mathbb{E}[aa^\top](x - x^*) + 2\mathbb{E}[\epsilon a^\top (x - x^*)] + \mathbb{E}[\epsilon]$
- $\mathbb{E}[aa^\top] = \text{Cov}(a) + \bar{a}\bar{a}^\top$
- SGD: $v = -\nabla_x (a_i^\top x - b_i)^2 = -2(a_i^\top x - b_i)a_i$
- Convex Hulls: Draw Graphs, Interior of polyhedral formed!
- Projections: Use KKT methods
- Epigraph of function: epi$f = \{(x,\mu) \mid f(x) \leq \mu\}$
- Lipschitz function: $\exists L \geq 0$ s.t. $\|f(x) - f(y)\| \leq M\|x - y\|, \forall x, y$
- Any set containing zero vector: Linearly dependent (fast check for regular points).
- Lagrangian multipliers: measure sensitivity (rate of change) of $f(x^*)$ due to a small change in constraint. If $g_i(x) = 0 \to g_i(x) + \delta_i = 0$; $h_j(x) \leq 0 \to h_j(x) + \epsilon_j \leq 0 \implies x^* \xrightarrow{\sim} x^* + \sum_{i=1}^{m} \lambda_i \delta_i + \sum_{j=1}^{p} \mu_j \epsilon_j$
- Lagrangian duals: Not guaranteed that both (P) and (D) will have a solution even if one does (unless Convex!).