Lagrangian Duality

제약이 있는 optimization 문제는 gradient가 0인 지점으로 solution 찾기 어려움

Constrained optimization problem

minimize $f(x)$ subject to $g(x) \le 0$ & $h(x) = 0$

Duality: solve an optimization problem → different optimization (variable)

Lagrangian: $\mathcal{L}(x, \lambda, \nu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \nu_j h_j(x)$

($\lambda_i \ge 0$, $g_i(x) \le 0$)

Lagrangian multipliers

$D(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu)$

optimal value의 하한 → 최대화

Convex optimization

Conjugate Prior

posterior를 closed form으로 구하기 위해

likelihood의 형태가 주어질 때 prior를 어떤 분포 family에서 고르면 posterior도 prior와 같은 분포 family에 놓임

$p(\theta|x) \propto p(x|\theta)p(\theta)$

e.g., normal normal normal

    - $p(y=1|x) = \sigma(w^\top x)$, $\sigma(a) = \frac{1}{1+e^{-a}}$
- likelihood (Bernoulli)
    - $p(y|x,w) = p^y(1-p)^{1-y}$ where $p = \sigma(w^\top x)$
    - $p(y|x,w) = \sigma(w^\top x)^y(1-\sigma(w^\top x))^{1-y}$
- MLE objective
    - maximize likelihood over data
    - equivalent to minimizing NLL (negative log-likelihood)

$$p = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$$

- NLL = binary cross-entropy
    - $-\log p(y|x,w) = -y \log p - (1-y)\log(1-p)$
    - equivalent to Binary Cross Entropy (BCE)
    - gradient: $\nabla_w \mathcal{L}(w) = X^\top(p-y)$
- training
    - no closed-form solution in general
    - optimize with gradient descent or variants (SGD, Adam)

Convexity

any local minimum = global minimum

- determines whether a problem has a single global minimum
    - a property of functions and optimization problems
- intuition
    - a function is convex if the line segment between any two points lies above the function
    - no local minima other than the global minimum

Convex Function

- definition

    - a function f is convex if for all x, y and $\lambda \in [0, 1]$:
        - $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ Jensen inequality
- geometric interpretation $f[\mathbb{E}(x)] \leq \mathbb{E}[f(x)]$
    - bowl-shaped surface
    - any descent direction eventually leads to the same minimum
        - any local minimum is a global minimum
    - gradient descent is guaranteed to converge (with proper step size)
- Hessian-based Characterization
    - for twice-differentiable functions
        - f is convex if its Hessian is positive semi-definite
            - Hessian ≥ 0 → convex
    - intuition
        - eigenvalues of the Hessian represent curvature
        - all eigenvalues ≥ 0 → no direction curves downward
    - example
        - in linear regression loss
            - Hessian = $2X^\top X$
                - $\nabla_w \mathcal{L} = -\frac{2}{N} X^\top (y - Xw)$$ linear regression이 convexity를 만족하는 이유
            - always positive semi-definite ↳ linear regression에서의 gradient

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_i^\top w)^2$$

$$\nabla_w = -\frac{2}{N} X^\top (y - Xw)$$

- why convexity matters in ML
    - guarantees a unique global optimum
        - avoids issues like local minima, saddle points, etc.
    - simplifies optimization and theoretical analysis
- v.s. non-convex problems
    - neural networks, matrix factorization, deep generative models, etc.
    - does not mean impossible
        - but optimization guarantees are weaker
    - convex ≠ easy

easy 보다는 guarantee → 오히려 linearity가 계산·해석적의 쉬움을 보임

        - large-scale convex problems can still be expensive

Lipschitz Continuity

- a notion of smoothness of a function
    - controls how fast a function can change
- definition
    - a function f is Lipschitz continuous if there exists $L \geq 0$ such that
        - $\|f(x) - f(y)\| \leq L\|x - y\|$ for all x, y
- intuition
    - the function has a bounded slope
    - L is the maximum rate of change
    - prevents the function from changing too abruptly

Lipschitz Continuous Gradient (Smoothness)

gradient의 변화가 bounded → 너무 급격하게 변하지 X

- a differentiable function f has an L-Lipschitz continuous gradient if
    - $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
- interpretation
    - the gradient does not change too fast
    - the curvature of the function is bounded
- equivalently
    - the Hessian eigenvalues are bounded:
        - $\nabla^2 f(x) \preceq LI$
- ㅈhy Lipschitz continuity matters in optimization
    - guarantees stability of gradient descent
    - allows choosing a safe learning rate
        - if gradient is L-Lipschitz:
            - step size $\eta \leq \frac{1}{L}$ ensures descent