# CS182 Introduction to Machine Learning, Fall 2023 Discussion3

Zhan-Wang Mao
maozhw@shanghaitech.edu.cn

November 2, 2023

# The Log-Sum-Exp Trick

- In statistical modeling and machine learning, we often work in a logarithmic scale
  - Multiplying small numbers

$$\log xy = \log x + \log y$$

  - Differentiation

$$\frac{\partial}{\partial x} \log[f(x)g(x)] = \frac{\partial}{\partial x} \log f(x) + \frac{\partial}{\partial x} \log g(x)$$

- Example: Multivariate Normal

$$\log(\det(\Sigma)) = \sum_{d=1}^{D} \log(\lambda_d)$$

https://gregorygundersen.com/blog/2020/02/09/log-sum-exp/

# The Log-Sum-Exp Trick

- Normalize an N-vector $x$ of log probabilities $x_i = \log p_i$

$$p_i = \frac{\exp(x_i)}{\sum_{n=1}^{N} \exp(x_n)}, \qquad \sum_{n=1}^{N} p_n = 1.$$

- Exponentiating might result in underflow or overflow
- Log-Sum-Exp operator

$$\mathrm{LSE}(x_1, \ldots, x_N) = \log\left(\sum_{n=1}^{N} \exp(x_n)\right).$$

# The Log-Sum-Exp Trick

- Perform the normalization using the Log-Sum-Exp operator

$$\exp(x_i) = p_i \sum_{n=1}^{N} \exp(x_n)$$

$$x_i = \log(p_i) + \log\left(\sum_{n=1}^{N} \exp(x_n)\right)$$

$$\log(p_i) = x_i - \log\left(\sum_{n=1}^{N} \exp(x_n)\right)$$

$$p_i = \exp\left(x_i - \underbrace{\log\sum_{n=1}^{N} \exp(x_n)}_{\mathrm{LSE}(x_1,\ldots,x_N)}\right).$$

- Does it really helps?

# The Log-Sum-Exp Trick

- Shift the values in the exponent by an arbitrary constant $C$

$$
\begin{aligned}
y &= \log\left(\sum_{n=1}^{N} \exp(x_n)\right) \\
e^y &= \sum_{n=1}^{N} \exp(x_n) \\
e^y &= e^c \sum_{n=1}^{N} \exp(x_n - c) \\
y &= c + \log \sum_{n=1}^{N} \exp(x_n - c).
\end{aligned}
$$

- Largest term as the reference $c = \max\{x_1, \ldots, x_N\}$

# Subgradients

- Recall that for convex and differentialble $f$,

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

Linear approximation always underestimates $f$.

- $g \in \mathbb{R}^n$ is a subgradient of convex function $f$ at $x$, if

$$f(y) \geq f(x) + g^T (y - x) \quad \text{for all } y$$

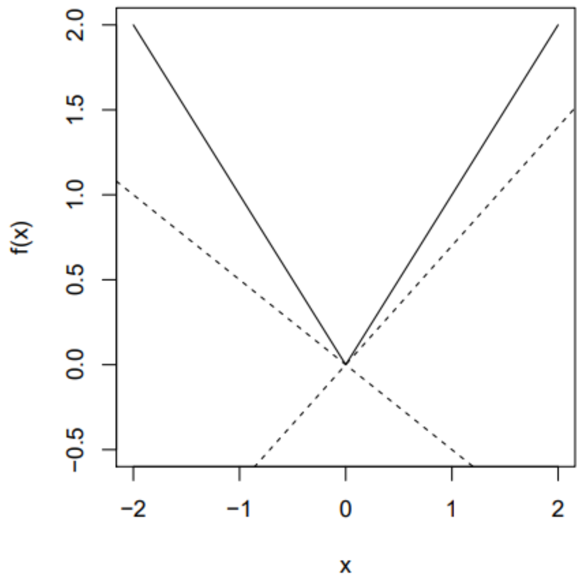
  - Always exists
  - If $f$ differentiable at $x$, then unique $g = \nabla f(x)$
  - Same definition works for nonconvex $f$ (however, subgradients need not exist)

- Subdifferential: $\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x\}$

https://www.stat.cmu.edu/~ryantibs/convexopt/lectures/subgrad.pdf

# Examples of Subgradients

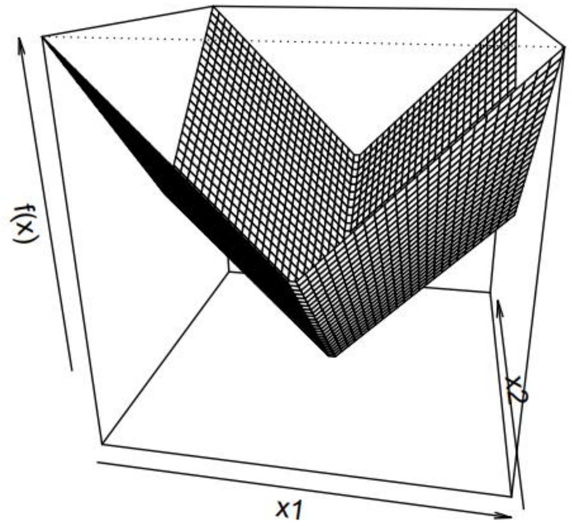
Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$




- For $x \neq 0$, unique subgradient $g = \text{sign}(x)$
- For $x = 0$, subgradient $g$ is any element of $[-1, 1]$

# Examples of Subgradients

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$




- For $x_i \neq 0$, unique $i$th component $g_i = \mathrm{sign}(x_i)$
- For $x_i = 0$, $i$th component $g_i$ is any element of $[-1, 1]$

# Optimality Condition

- Subgradient Optimality Condition: for any $f$ (convex or not)

$$f(x^\star) = \min_x f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x^\star)$$

- 0 is a subgradient of $f$ at $x^\star$, then for all $y$

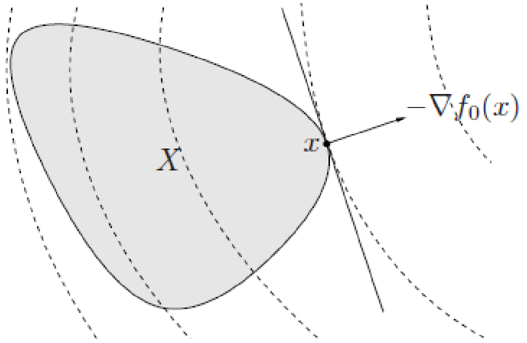
$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star)$$

- Recall first-order optimality condition: for convex problem

$$\min_x f(x) \quad \text{subject to} \quad x \in C$$

$f$ differentiable, a feasible point $x$ is optimal if and only if

$$\nabla f(x)^T(y - x) \geq 0 \quad \text{for all } y \in C$$




  - If $C = \mathbb{R}^n$ reduces to familiar $\nabla f(x) = 0$

# Lasso Optimality

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, lasso problem can be parametrized as

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

where $\lambda \geq 0$. Subgradient optimality:

$$\begin{aligned}
0 \in \partial\Big(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\Big) \\
&\iff \quad 0 \in -X^T(y - X\beta) + \lambda\partial\|\beta\|_1 \\
&\iff \quad X^T(y - X\beta) = \lambda v
\end{aligned}$$

for some $v \in \partial\|\beta\|_1$, i.e.,

$$v_i \in \begin{cases} \{1\} & \text{if } \beta_i > 0 \\ \{-1\} & \text{if } \beta_i < 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}, \quad i = 1, \ldots, p$$

# Lasso Solution*

Write $X_1, \ldots, X_p$ for columns of $X$. Then our condition reads:

$$\begin{cases} X_i^T(y - X\beta) = \lambda \cdot \mathrm{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ |X_i^T(y - X\beta)| \leq \lambda & \text{if } \beta_i = 0 \end{cases}$$

Simplfied lasso problem with $X = I$:

$$\min_{\beta} \frac{1}{2}\|y - \beta\|_2^2 + \lambda\|\beta\|_1$$

This we can solve directly using subgradient optimality. Solution is $\beta = S_\lambda(y)$, where $S_\lambda$ is the soft-thresholding operator:

# Lasso Solution*

- **Solution:** $\beta = S_\lambda(y)$, where $S_\lambda$ is the soft-thresholding operator:

$$\beta_i = [S_\lambda(y)]_i = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } -\lambda \le y_i \le \lambda \\ y_i + \lambda & \text{if } y_i < -\lambda \end{cases}, \quad i = 1, \dots, n$$

Check: from last slide, subgradient optimality conditions are

$$\begin{cases} y_i - \beta_i = \lambda \cdot \mathrm{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ |y_i - \beta_i| \le \lambda & \text{if } \beta_i = 0 \end{cases}$$

- $X \neq I$? Proximal Gradient Method!

# Lagrangian

- **standard form problem** (not necessarily convex)

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

  variable $x \in \mathbf{R}^n$, domain $\mathcal{D}$, optimal value $p^\star$

- **Lagrangian:** $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$, with $\mathbf{dom}\, L = \mathcal{D} \times \mathbf{R}^m \times \mathbf{R}^p$,

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

  - weighted sum of objective and constraint functions
  - $\lambda_i$ is **Lagrange multiplier** associated with $f_i(x) \leq 0$
  - $\nu_i$ is Lagrange multiplier associated with $h_i(x) = 0$

https://web.stanford.edu/class/ee364a/lectures/duality.pdf

# Lagrange Dual Function

- **Lagrange dual function:** $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$,

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left( f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)$$

- $g$ is concave, can be $-\infty$ for some $\lambda$, $\nu$
- **lower bound property:** if $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^\star$
- proof: if $\tilde{x}$ is feasible and $\lambda \succeq 0$, then

$$f_0(\tilde{x}) \geq L(\tilde{x}, \lambda, \nu) \geq \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = g(\lambda, \nu)$$

  minimizing over all feasible $\tilde{x}$ gives $p^\star \geq g(\lambda, \nu)$

# The Lagrange Dual Problem

(Lagrange) **dual problem**

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

- finds best lower bound on $p^\star$, obtained from Lagrange dual function
- a convex optimization problem, even if original **primal** problem is not
- dual optimal value denoted $d^\star$
- $\lambda$, $\nu$ are dual feasible if $\lambda \succeq 0$, $(\lambda, \nu) \in \mathbf{dom}\, g$
- often simplified by making implicit constraint $(\lambda, \nu) \in \mathbf{dom}\, g$ explicit

# Weak and Strong Duality

- Weak Duality: $d^\star \leq p^\star$
  - Always holds
- Strong Duality: $d^\star = p^\star$
  - Does not hold in general
  - (Usually) hold for convex problems
  - Conditions that guarantee strong duality in convex problems are called constraint qualifications
- Slater's Condition: strong duality hold for a convex problem if it is strictly feasible, i.e.,

$$\text{there is an } x \in \textbf{int}\, \mathcal{D} \text{ with } f_i(x) < 0,\ i = 1, \ldots, m,\ Ax = b$$

# Complementary Slackness

- assume strong duality holds, $x^\star$ is primal optimal, $(\lambda^\star, \nu^\star)$ is dual optimal

$$
\begin{aligned}
f_0(x^\star) = g(\lambda^\star, \nu^\star) &= \inf_x \left( f_0(x) + \sum_{i=1}^m \lambda_i^\star f_i(x) + \sum_{i=1}^p \nu_i^\star h_i(x) \right) \\
&\le f_0(x^\star) + \sum_{i=1}^m \lambda_i^\star f_i(x^\star) + \sum_{i=1}^p \nu_i^\star h_i(x^\star) \\
&\le f_0(x^\star)
\end{aligned}
$$

- hence, the two inequalities hold with equality
- $x^\star$ minimizes $L(x, \lambda^\star, \nu^\star)$
- $\lambda_i^\star f_i(x^\star) = 0$ for $i = 1, \ldots, m$ (known as **complementary slackness**):

$$
\lambda_i^\star > 0 \implies f_i(x^\star) = 0, \qquad f_i(x^\star) < 0 \implies \lambda_i^\star = 0
$$

# Karush-Kuhn-Tucker (KKT) Conditions

- Make no assumptions about convexity
- If strong duality holds,

$$x^\star \text{ and } (\lambda^\star, \nu^\star) \text{ be any primal and dual optimal points}$$

- Thus, they must satisfy
  - Primal feasible: $$\begin{aligned} f_i(x^\star) &\leq 0, \quad i = 1, \ldots, m \\ h_i(x^\star) &= 0, \quad i = 1, \ldots, p \end{aligned}$$
  - Dual feasible: $\lambda_i^\star \geq 0, \quad i = 1, \ldots, m$
  - Complementary Slackness: $\lambda_i^\star f_i(x^\star) = 0, \quad i = 1, \ldots, m$
  - Gradient of Lagrangian with respect to *x* vanishes:

$$\nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{i=1}^{p} \nu_i^\star \nabla h_i(x^\star) = 0.$$

# KKT Conditions for Convex Problems

- When the primal problem is convex, KKT conditions are also sufficient
- Assume $\tilde{x}, \tilde{\lambda}, \tilde{\nu}$ are any points that satisfy the KKT conditions

$$\begin{aligned} f_i(\tilde{x}) &\leq 0, \quad i = 1, \ldots, m \\ h_i(\tilde{x}) &= 0, \quad i = 1, \ldots, p \\ \tilde{\lambda}_i &\geq 0, \quad i = 1, \ldots, m \\ \tilde{\lambda}_i f_i(\tilde{x}) &= 0, \quad i = 1, \ldots, m \\ \nabla f_0(\tilde{x}) + \sum_{i=1}^m \tilde{\lambda}_i \nabla f_i(\tilde{x}) + \sum_{i=1}^p \tilde{\nu}_i \nabla h_i(\tilde{x}) &= 0, \end{aligned}$$

Then $\tilde{x}$ and $(\tilde{\lambda}, \tilde{\nu})$ are primal and dual optimal, with zero duality gap. $\Leftrightarrow$ KKT conditions

- From Complementary Slackness: $f_0(\tilde{x}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$
- From last condition and convexity: $g(\tilde{\lambda}, \tilde{\nu}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$
- If Slater's condition is satisfied, then

*$x$ is optimal if and only if there exist $\lambda$, $\nu$ that satisfy KKT conditions*

# Constrained and Lagrange Forms

Often in statistics and machine learning we'll switch back and forth between constrained form, where $t \in \mathbb{R}$ is a tuning parameter,

$$\min_x \; f(x) \quad \text{subject to} \quad h(x) \leq t \tag{C}$$

and Lagrange form, where $\lambda \geq 0$ is a tuning parameter,

$$\min_x \; f(x) + \lambda \cdot h(x) \tag{L}$$

and claim these are equivalent. Is this true (assuming convex $f, h$)?

(C) to (L): if (C) is strictly feasible, then strong duality holds, and there exists $\lambda \geq 0$ (dual solution) such that any solution $x^\star$ in (C) minimizes

$$f(x) + \lambda \cdot (h(x) - t)$$

so $x^\star$ is also a solution in (L)

(L) to (C): if $x^\star$ is a solution in (L), then the KKT conditions for (C) are satisfied by taking $t = h(x^\star)$, so $x^\star$ is a solution in (C)