CS 446/ECE 449: Machine Learning

A. G. Schwing

University of Illinois at Urbana-Champaign, 2020

Scribe & Exercises

L5: Optimization Dual

Goals of this lecture

- Constrained optimization
- Understanding duality for optimization

Reading Material

- S. Boyd and L. Vandenberghe; Convex Optimization; Chapter 5

Optimization problems that we have seen so far:

- Linear Regression

$$\min_{\mathbf{w}} \frac{1}{2} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left(y^{(i)} - \phi(x^{(i)})^\top \mathbf{w}\right)^2$$

- Logistic Regression

$$\min_{\mathbf{w}} \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log\left(1 + \exp(-y^{(i)}\mathbf{w}^T\phi(x^{(i)}))\right)$$

Finding optimum:

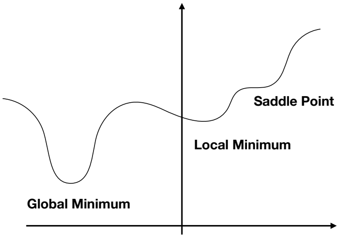
Analytically computable optimum vs. gradient descent

The Problem more generally:

$$\begin{array}{ll} \min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) \\ \text{s.t.} & f_i(\boldsymbol{w}) \leq 0 \quad \forall i \in \{1, \ldots, C\} \end{array}$$

Solution:

Solution $\boldsymbol{w}^*$ has smallest value $f_0(\boldsymbol{w}^*)$ among all values that satisfy constraints



Original/Primal Problem:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & f_0(\boldsymbol{w}) \\ \text{s.t.} \quad & f_i(\boldsymbol{w}) \leq 0 \quad \forall i \in \{1, \ldots, C_1\} \\ & h_i(\boldsymbol{w}) = 0 \quad \forall i \in \{1, \ldots, C_2\} \end{aligned}$$

How to optimize this?

Original/Primal Problem:

$$\begin{array}{lll} \min_{\boldsymbol{w}} & f_0(\boldsymbol{w}) & \\ \text{s.t.} & f_i(\boldsymbol{w}) \leq 0 & \forall i \in \{1, \ldots, C_1\} \\ & h_i(\boldsymbol{w}) = 0 & \forall i \in \{1, \ldots, C_2\} \end{array}$$

Lagrangian

$$L(\boldsymbol{w}, \lambda, \nu) = f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

- λ_i are Lagrange multiplier associated with inequality constraints
- ν_i are Lagrange multiplier associated with equality constraints

Properties of Lagrangian:

$$L(\boldsymbol{w}, \lambda, \nu) = f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

If $\hat{\boldsymbol{w}}$ feasible and $\lambda_i \geq 0$ $\forall i$ then

$$f_0(\hat{\boldsymbol{w}}) \geq L(\hat{\boldsymbol{w}}, \lambda, \nu) \geq \min_{\boldsymbol{w}\in\mathcal{W}} L(\boldsymbol{w}, \lambda, \nu) = g(\lambda, \nu) \quad \forall \lambda \geq 0, \nu$$

$$f_0(\boldsymbol{w}^*) \geq g(\lambda, \nu) \quad \forall \lambda \geq 0, \nu$$

$\mathcal{W}$ denotes all the constraints that are not part of the Lagrangian (larger than feasible set)

Dual Program:

$$\max_{\lambda,\nu} g(\lambda, \nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

Recipe for computing dual program:

- Bring primal program into standard form
- Assign Lagrange multipliers to a suitable set of constraints
- Subsume all other constrains in $\mathcal{W}$
- Write down the Lagrangian L
- Minimize Lagrangian w.r.t. primal variables s.t. $\boldsymbol{w} \in \mathcal{W}$

Examples: Linear Program

$$\min_{\boldsymbol{w}} \boldsymbol{c}^\top \boldsymbol{w} \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{w} \leq \mathbf{b}$$

Lagrangian: $(\lambda \geq 0)$

$$L() = \boldsymbol{c}^\top \boldsymbol{w} + \lambda^\top(\boldsymbol{A}\boldsymbol{w} - \mathbf{b}) = (c + \boldsymbol{A}^\top\lambda)^\top \boldsymbol{w} - \mathbf{b}^\top\lambda$$

Minimizing Lagrangian w.r.t. primal variables:

$$\min_{\boldsymbol{w}} L() = \begin{cases} -\mathbf{b}^\top\lambda & \boldsymbol{A}^\top\lambda + \boldsymbol{c} = 0 \\ -\infty & \text{otherwise} \end{cases}$$

Dual Program:

$$\max_{\lambda \geq 0} -\mathbf{b}^\top\lambda \quad \text{s.t.} \quad \boldsymbol{A}^\top\lambda + \boldsymbol{c} = 0,$$

Examples: Logistic Regression

no constraint -> come up with one

$$\min_{\boldsymbol{w}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})))$$

Reformulate:

$$\min_{\boldsymbol{w},z^{(i)}} \frac{C}{2}\|\boldsymbol{w}\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log(1+\exp(-z^{(i)})) \quad \text{s.t.} \quad \lambda \quad z^{(i)} = y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)})$$

Lagrangian:

$$\begin{aligned} L() &= \frac{C}{2}\|\boldsymbol{w}\|_2^2 - \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)}y^{(i)}\boldsymbol{w}^\top\phi(x^{(i)}) \\ &+ \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \left[\log(1+\exp(-z^{(i)})) + \lambda^{(i)}z^{(i)}\right] \end{aligned}$$

Minimize Lagrangian w.r.t. primal variables ($\min_{\boldsymbol{w},z} L()$):

$$\frac{\partial L}{\partial \boldsymbol{w}}: \qquad C\boldsymbol{w} = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})$$

$$\frac{\partial L}{\partial z^{(i)}}: \qquad \lambda^{(i)} = \frac{\exp(-z^{(i)})}{1+\exp(-z^{(i)})} \qquad \Longrightarrow \lambda^{(i)} \geq 0$$

$$\Longrightarrow z^{(i)} = \log \frac{1-\lambda^{(i)}}{\lambda^{(i)}} \qquad \Longrightarrow \lambda^{(i)} \leq 1$$

Dual function:

$$g(\lambda) = -\frac{1}{2C}\| \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \lambda^{(i)} y^{(i)} \phi(x^{(i)})\|_2^2 + \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} H(\lambda^{(i)})$$

with binary entropy $H(\lambda^{(i)})$

Dual program:

$$\max_{\lambda} g(\lambda) \quad \text{s.t.} \quad 0 \leq \lambda^{(i)} \leq 1 \qquad \forall (x^{(i)}, y^{(i)}) \in \mathcal{D}$$

Instead of optimizing the primal we can optimize the dual and convert the result

Why is this useful?

- Sometimes less constraints
- Sometimes easier to optimize
- Sometimes interesting insights
- Sometimes lower bounds

Properties of Dual Program

$$\max_{\lambda,\nu} g(\lambda,\nu) \quad \text{s.t.} \quad \lambda_i \geq 0 \quad \forall i$$

- May only have simple constraints if at all
- Can be used for sensitivity analysis
- Lower-bounds the optimal primal value
- Dual Program is always concave:

$$g(\lambda,\nu) = \min_{\boldsymbol{w}\in\mathcal{W}} L(\boldsymbol{w},\lambda,\nu) := f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i h_i(\boldsymbol{w})$$

  - Pointwise minimum
  - Affine functions in λ,ν

Weak duality:

$$f(\boldsymbol{w}^*) \geq g(\lambda^*, \nu^*)$$

- Always holds (for convex and non-convex problems)
- Can be used to find nontrivial lower bounds

Strong duality:

$$f(\boldsymbol{w}^*) = g(\lambda^*, \nu^*)$$

- Does not hold in general
- (Usually) holds for convex problems
- Conditions that guarantee strong duality in convex problems are called **constraint qualifications**

Consequence of strong duality:

Assume strong duality holds:

$$f_0(\boldsymbol{w}^*) = g(\lambda^*, \nu^*) \quad = \quad \min_{\boldsymbol{w}\in\mathcal{W}} \left(f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i^* f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i^* h_i(\boldsymbol{w}) \right)$$

$$= \quad f_0(\boldsymbol{w}^*) + \sum_{i=1}^{C_1} \lambda_i^* f_i(\boldsymbol{w}^*) + \sum_{i=1}^{C_2} \nu_i^* h_i(\boldsymbol{w}^*)$$

Consequently:

$$\lambda_i^* f_i(\boldsymbol{w}^*) = 0 \quad \forall i \in \{1, \ldots, C_1\}$$

$$\lambda_i^* > 0 \implies f_i(\boldsymbol{w}^*) = 0, \qquad f_i(\boldsymbol{w}^*) < 0 \implies \lambda_i^* = 0$$

known as **complementary slackness**

Karush-Kuhn-Tucker (KKT) conditions

- Primal feasibility: $f_i(\boldsymbol{w}) \leq 0 \quad \forall i \in \{1, \ldots, C_1\}$;
  $h_i(\boldsymbol{w}) = 0 \quad \forall i \in \{1, \ldots, C_2\}$
- Dual feasibility: $\lambda_i \geq 0 \quad \forall i \in \{1, \ldots, C_1\}$
- Complementary slackness: $\lambda_i f_i(\boldsymbol{w}) = 0 \quad \forall i \in \{1, \ldots, C_1\}$
- Stationarity of Lagrangian:

$$\nabla f_0(\boldsymbol{w}) + \sum_{i=1}^{C_1} \lambda_i \nabla f_i(\boldsymbol{w}) + \sum_{i=1}^{C_2} \nu_i \nabla h_i(\boldsymbol{w}) = 0$$

If strong duality holds and $\boldsymbol{w}, \lambda, \nu$ are optimal, then they must satisfy the KKT conditions

Converse is true for convex problems, i.e., if $\boldsymbol{w}, \lambda, \nu$ satisfy KKT conditions, then they are optimal

Quiz:

- What to do before computing the Lagrangian?
- How to obtain the dual program?
- Why duality?

Important topics of this lecture

- Lagrangian
- Dual program

Up next:

- Support vector machines