Vũ Hữu Tiệp

Convex Optimization - Tối Ưu Lồi

Theo blog: http://machinelearningcoban.com

(*Đang trong quá trình xây dựng*)

First Edition

June 24, 2017

Contents

18 Duality 1

18

Duality

Trong bài viết này, chúng ta giả sử rằng các đạo hàm đều tồn tại.

Bài viết này chủ yếu được dịch lại từ Chương 5 của cuốn *Convex Optimization* trong tài liệu tham khảo.

18.1 Giới thiệu

Trong Bài 16, chúng ta đã làm quen với các khái niệm về tập hợp lồi và hàm số lồi. Tiếp theo đó, trong Bài 17, tôi cũng đã trình bày về các bài toán tối ưu lồi, cách nhận dạng và cách sử dụng thư viện để giải các bài toán lồi cơ bản. Trong bài này, chúng ta sẽ tiếp tục tiếp cận một cách sâu hơn: các điều kiện về nghiệm của các bài toán tối ưu, cả lồi và không lồi; bài toán đối ngẫu (dual problem) và điều kiện KKT.

Trước tiên, chúng ta lại bắt đầu bằng những kỹ thuật đơn giản cho các bài toán cơ bản. Kỹ thuật này có lẽ các bạn đã từng nghe đến: Phương pháp nhân tử Lagrange (method of Lagrange multipliers). Đây là một phương pháp giúp tìm các điểm cực trị của hàm mục tiêu trên feasible set của bài toán.

Nhắc lại rằng giá trị lớn nhất và nhỏ nhất (nếu có) của một hàm số $f_0(\mathbf{x})$ khả vi (và tập xác định là một *tập mở*) đạt được tại một trong các điểm cực trị của nó. Và điều kiện cần để một điểm là điểm cực trị là đạo hàm của hàm số tại điểm này $f_0'(x) = 0$. Chú ý rằng một điểm thoả mãn $f_0'(\mathbf{x}) = 0$ thì được gọi là *điểm dừng* hay *stationary point*. Điểm cực trị là một điểm dừng nhưng không phải điểm dừng nào cũng là điểm cực trị. Ví dụ hàm $f(x) = x^3$ có 0 là một điểm dừng nhưng không phải là điểm cực trị.

Với hàm nhiều biến, ta cũng có thể áp dụng quan sát này. Tức chúng ta cần đi tìm nghiệm của phương trình đạo hàm *theo mỗi biến* bằng 0. Tuy nhiên, đó là với các bài toán không ràng buộc (unconstrained optimization problems), với các bài toán có ràng buộc như chúng ta đã gặp trong Bài 17 thì sao?

Trước tiên chúng ta xét bài toán mà ràng buộc chỉ là một phương trình:

$$\mathbf{x} = \arg\min_{\mathbf{x}} f_0(\mathbf{x}) \tag{18.1}$$
$$\text{subject to: } f_1(\mathbf{x}) = 0 \quad (1) \tag{18.2}$$

Bài toán này là bài toán tổng quát, không nhất thiết phải lồi. Tức hàm mục tiêu và hàm ràng buộc không nhất thiết phải lồi.

18.2 Phương pháp nhân tử Lagrange

Nếu chúng ta đưa được bài toán này về một bài toán không ràng buộc thì chúng ta có thể tìm được nghiệm bằng cách giải hệ phương trình đạo hàm theo từng thành phần bằng 0 (giả sử rằng việc giải hệ phương trình này là khả thi).

Điều này là động lực để nhà toán học Lagrange sử dụng hàm số: $\mathcal{L}(\mathbf{x}, \lambda) = f_0(\mathbf{x}) + \lambda f_1(\mathbf{x})$. Chú ý rằng, trong hàm số này, chúng ta có thêm một biến nữa là λ, biến này được gọi là nhân tử Lagrange (Lagrange multiplier). Hàm số $\mathcal{L}(\mathbf{x}, \lambda)$ được gọi là *hàm hỗ trợ* (*auxiliary function*), hay *the Lagrangian*. Người ta đã chứng minh được rằng, điểm *optimal value* của bài toán (1) thoả mãn điều kiện $\nabla_{\mathbf{x},\lambda}\mathcal{L}(\mathbf{x}, \lambda) = 0$ (tôi xin được bỏ qua chứng minh của phần này). Điều này tương dương với:

$$\nabla_{\mathbf{x}} f_0(\mathbf{x}) + \lambda \nabla_{\mathbf{x}} f_1(\mathbf{x}) = 0 \quad (2) \tag{18.3}$$
$$f_1(\mathbf{x}) = 0 \quad (3) \tag{18.4}$$

Để ý rằng điều kiện thứ hai chính là $\nabla_\lambda \mathcal{L}(\mathbf{x}, \lambda) = 0$, và cũng chính là ràng buộc trong bài toán (1).

Việc giải hệ phương trình (2) – (3), trong nhiều trường hợp, đơn giản hơn việc trực tiếp đi tìm *optimal value* của bài toán (1).

Xét các ví dụ đơn giản sau đây.

18.2.1 Ví dụ

Ví dụ 1: Tìm giá trị lớn nhất và nhỏ nhất của hàm số $f_0(x, y) = x + y$ thoả mãn điều kiện $f_1(x, y) = x^2 + y^2 = 2$. Ta nhận thấy rằng đây không phải là một bài toán tối ưu lồi vì *feasible set* $x^2 + y^2 = 2$ không phải là một tập lồi (nó chỉ là một đường tròn).

Lời giải:

Lagrangian của bài toán này là: $\mathcal{L}(x, y, \lambda) = x + y + \lambda(x^2 + y^2 - 2)$. Các điểm cực trị của hàm số Lagrange phải thoả mãn điều kiện:

$$\nabla_{x,y,\lambda}\mathcal{L}(x,y,\lambda)=0 \Leftrightarrow \begin{cases} 1+2\lambda x &= 0 \quad (4)\\ 1+2\lambda y &= 0 \quad (5)\\ x^2+y^2 &= 2 \quad (6)\end{cases}$$

Từ (4) và (5) ta suy ra $x=y=\frac{-1}{2\lambda}$. Thay vào (6) ta sẽ có $\lambda^2=\frac{1}{4}\Rightarrow\lambda=\pm\frac{1}{2}$. Vậy ta được 2 cặp nghiệm $(x,y)\in$
$(1,1),(-1,-1)$. Bằng cách thay các giá trị này vào hàm mục tiêu, ta tìm được giá trị nhỏ nhất và lớn nhất của hàm số cần tìm.

Ví dụ 2: Cross-entropy. Trong bài Bài 10 và Bải 13, chúng ta đã được biết đến hàm mất mát ở dạng *cross entropy*. Chúng ta cũng đã biết rằng hàm cross entropy được dùng để đo sự giống nhau của hai phân phối xác suất với giá trị của hàm số này càng nhỏ thì hai xác suất càng gần nhau. Chúng ta cũng đã phát biểu rằng giá trị nhỏ nhất của hàm cross entopy đạt được khi từng gặp xác suất là giống nhau. Bây giờ, tôi xin phát biểu lại và chứng minh nhận định trên.

Cho một phân bố xác xuất $\mathbf{p}=[p_1,p_2,\ldots,p_n]^T$ với $p_i\in[0,1]$ và $\sum_{i=1}^n p_i=1$. Với một phân bố xác suất bất kỳ $\mathbf{q}=[q_1,q_2,\ldots,q_n]$ và giả sử rằng $q_i\neq 0,\forall i$, hàm số cross entropy được định nghĩa là:

$$f_0(\mathbf{q})=-\sum_{i=1}^n p_i\log(q_i)$$

Hãy tìm $\mathbf{q}$ để hàm cross entropy đạt giá trị nhỏ nhất.

Trong bài toán này, ta có ràng buộc là $\sum_{i=1}^n q_i=1$. *Lagrangian* của bài toán là:

$$\mathcal{L}(q_1,q_2,\ldots,q_n,\lambda)=-\sum_{i=1}^n p_i\log(q_i)+\lambda(\sum_{i=1}^n q_i-1)$$

Ta cần giải hệ phương trình:

$$\nabla_{q_1,\ldots,q_n,\lambda}\mathcal{L}(q_1,\ldots,q_n,\lambda)=0 \Leftrightarrow \begin{cases} -\frac{p_i}{q_i}+\lambda &= 0, \quad i=1,\ldots,n \quad (7)\\ q_1+q_2+\cdots+q_n &= 1 \quad (8)\end{cases}$$

Từ (7) ta có $p_i=\lambda q_i$. Vậy nên: $1=\sum_{i=1}^n p_i=\lambda\sum_{i=1}^n q_i=\lambda\Rightarrow\lambda=1\Rightarrow q_i=p_i,\forall i$.

Qua đây, chúng ta đã hiểu rằng vì sao hàm số cross entropy được dùng để *ép* hai xác suất *gần nhau*.

18.3 Hàm đối ngẫu Lagrange (The Lagrange dual function)

18.3.1 Lagrangian

Với bài toán tối ưu tổng quát:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f_0(\mathbf{x}) \tag{18.5}$$

$$\text{subject to:} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m \quad (9) \tag{18.6}$$

$$h_j(\mathbf{x}) = 0, \quad j = 1, 2, \ldots, p \tag{18.7}$$

với miền xác đinh $\mathcal{D} = (\cap_{i=0}^{m} \text{dom} f_i) \cap (\cap_{j=1}^{p} \text{dom} h_j)$. Chú ý rằng, chúng ta đang không giả sử về tính chất lồi của hàm tối ưu hay các hàm ràng buộc ở đây. Giả sử duy nhất ở đây là $\mathcal{D} \neq \emptyset$ (tập rỗng).

Lagrangian cũng được xây dựng tương tự với mỗi nhân tử Lagrange cho một (bất) phương trình ràng buộc:

$$\mathcal{L}(\mathbf{x}, \lambda, \nu) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{p} \nu_j h_j(\mathbf{x})$$

với $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_m]; \nu = [\nu_1, \nu_2, \ldots, \nu_p]$ (*ký hiệu ν này không phải là chữ v mà là chữ nu trong tiếng Hy Lạp, đọc như từ new*) là các vectors và được gọi là *dual variables* (*biến đối ngẫu*) hoặc *Lagrange multiplier vectors* (vector nhân tử Lagrange). Lúc này nếu biến chính $\mathbf{x} \in \mathbb{R}^n$ thì tổng số biến của hàm số này sẽ là $n + m + p$.

(*Thông thường, tôi dùng các chữ cái viết thường in đậm để biểu diễn một vector, trong trường hợp này tôi không bôi đậm được λ và ν do hạn chế của LaTeX khi viết cùng markdown. Tôi lưu ý điều này để hạn chế nhầm lẫn cho bạn đọc*)

18.3.2 Hàm đối ngẫu Lagrange

Hàm đối ngẫu Lagrange của bài toán tối ưu (hoặc gọn là *hàm số đối ngẫu*) (9) là một hàm của các biến đối ngẫu, được định nghĩa là giá trị nhỏ nhất theo $\mathbf{x}$ của *Lagrangian*:

$$g(\lambda, \nu) = \inf_{\mathbf{x} \in \mathcal{D}} \mathcal{L}(\mathbf{x}, \lambda, \nu) \tag{18.8}$$

$$= \inf_{\mathbf{x} \in \mathcal{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{p} \nu_j h_j(\mathbf{x}) \right) \tag{18.9}$$

Nếu *Lagrangian không bị chặn dưới*, hàm đối ngẫu tại λ, ν sẽ lấy giá trị $-\infty$.

Đặc biệt quan trọng:

- inf được lấy trên miền $x \in \mathcal{D}$, tức miền xác định của bài toán (là giao của miền xác định của mọi hàm trong bài toán). Miền xác định này khác với *feasible set*. Thông thường, *feasible set* là tập con của miền xác định $\mathcal{D}$.

- Với mỗi $\mathbf{x}$, *Lagrangian* là một hàm *affine* của (λ, ν), tức là một hàm *concave*. Vậy, *hàm đối ngẫu* chính là *pointwise infimum* của (có thể vô hạn) các hàm concave, tức là một hàm concave. Vậy **hàm đối ngẫu của một bài toán tối ưu bất kỳ là một hàm concave, bất kể bài toán ban đầu có phải là convex hay không**. Nhắc lại rằng *pointwise supremum* của các hàm *convex* là một hàm *convex*, và một hàm là *concave* nếu khi đổi dấu hàm đó, ta được một hàm *convex*.

18.3.3 Chặn dưới của giá trị tối ưu

Nếu p^* là *optimal value* (giá trị tối ưu) của bài toán (9), thì với các biến đối ngẫu $\lambda_i \geq 0, \forall i$ và ν *bất kỳ*, chúng ta sẽ có:

$$g(\lambda, \nu) \leq p^* \quad (10)$$

Tính chất này có thể được chứng minh dễ dàng. Giả sử $\mathbf{x}_0$ là một điểm *feasible* bất kỳ của bài toán (9), tức thoả mãn các điều kiện ràng buộc $f_i(\mathbf{x}_0) \leq 0, \forall i = 1, \ldots, m; h_j(\mathbf{x}_0) = 0, \forall j = 1, \ldots, p$, ta sẽ có:

$$\sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}_0) + \sum_{j=1}^{p} \nu_j h_j(\mathbf{x}_0) \leq 0 \Rightarrow \mathcal{L}(\mathbf{x}_0, \lambda, \nu) \leq f_0(\mathbf{x}_0)$$

Vì điều này đúng với mọi $\mathbf{x}_0$ *feasible*, ta sẽ có tính chất quan trọng sau đây:

$$g(\lambda, \nu) = \inf_{\mathbf{x} \in \mathcal{D}} \mathcal{L}(\mathbf{x}, \lambda, \nu) \leq \mathcal{L}(\mathbf{x}_0, \lambda, \nu) \leq f_0(\mathbf{x}_0).$$

Khi $\mathbf{x}_0 = \mathbf{x}^*$, ta có bất đẳng thức (10).

18.3.4 Ví dụ

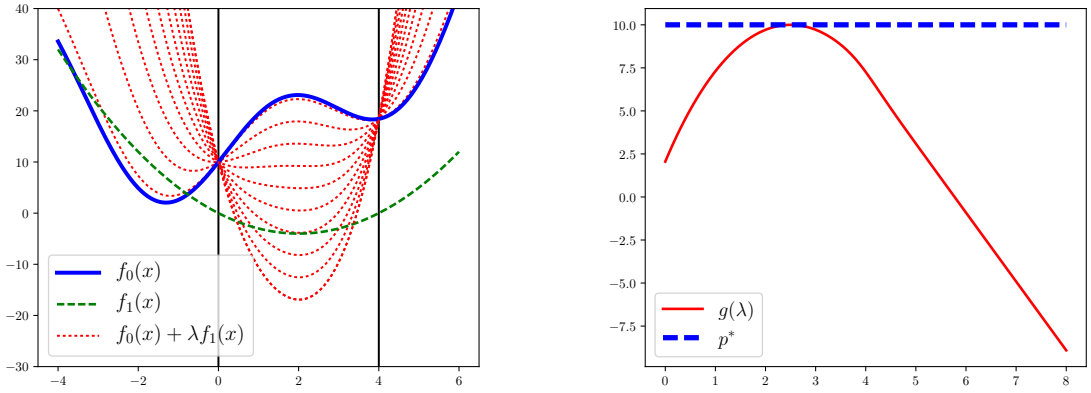
Ví dụ 1

Xét bài toán tối ưu sau:

$$x = \arg\min_x x^2 + 10\sin(x) + 10 \quad (18.10)$$

$$\text{subject to:} \quad (x-2)^2 \leq 4 \quad (18.11)$$

Chú ý: Với bài toán này, miền xác định $\mathcal{D} = \mathbb{R}$ nhưng *feasible set* là $0 \leq x \leq 4$.



Hình 18.1: Ví dụ về dual function. Trái: Đường màu lam đậm thể hiện hàm mục tiêu. Đường nét đứt mà lục thể hiện hàm số ràng buộc. Các đường nét đứt màu đỏ thể hiện dual function ứng với các λ khác nhau. Phải: Đường nét đứt thể hiện giá trị tối ưu của bài toán . Đường màu đỏ thể hiện dual function. Với mọi λ, giá trị của hàm dual function nhỏ hơn hoặc bằng giá trị tối ưu của bài toán gốc.

Với hàm mục tiêu là đường đậm màu xanh lam trong Hình 18.1. Ràng buộc thực ra $0 \leq x \leq 4$, nhưng tôi viết ở dạng này để bài toán thêm phần thú vị. Hàm số ràng buộc $f_1(x) = (x-2)^2 - 4$ được cho bởi đường nét đứt màu xanh lục. Optimal value của bài toán này có thể được nhận ra là điểm trên đồ thị có hoành độ bằng 0. Chú ý rằng hàm mục tiêu ở đây không phải là hàm lồi nên bài toán tối ưu này cũng không phải là lồi, mặc dù hàm bất phương trình ràng buộc $f_1(x)$ là lồi.

Lagrangian của bài toàn này có dạng:

$$\mathcal{L}(x, \lambda) = x^2 + 10\sin(x) + 10 + \lambda((x-2)^2 - 4)$$

Các đường dấu chấm màu đỏ trong Hình 1 là các đường ứng với các λ khác nhau. Vùng bị chặn giữa hai đường thẳng đứng màu đen thể hiện miền *feasible* của bài toán tối ưu.

Với mỗi λ, *dual function* được định nghĩa là:

$$g(\lambda) = \inf_x \left(x^2 + 10\sin(x) + 10 + \lambda((x-2)^2 - 4)\right), \quad \lambda \geq 0.$$

Từ hình 1 bên trái, ta có thể thấy ngay rằng với các λ khác nhau, $g(\lambda)$ hoặc tại điểm có hoành độ bằng 0, hoặc tại một điểm thấp hơn điểm tối ưu của bài toán. Đồ thị của hàm $g(\lambda)$ được cho bởi đường liền màu đỏ ở Hình 1 bên phải. Đường nét đứt màu lam thể hiện *optimal value* của bài toán tối ưu ban đầu. Ta có thể thấy ngay hai điều:

- Đường liền màu đỏ luôn nằm dưới (hoặc có đoạn trùng) với đường nét đứt màu lam.

- Hàm $g(\lambda)$ có dạng một hàm *concave*, tức nếu ta *lật* đồ thị này theo hướng trên-dưới thì ta sẽ có đồ thị của một hàm *convex*. (Mặc dù bài toán tối ưu gốc là không phải là một bài toán lồi.)

(*Để vẽ được hình bên phải, tôi đã dùng Gradient Descent để tìm giá trị nhỏ nhất ứng với mỗi* λ.)

Ví dụ 2

Xét một bài toán Linear Programming:

$$x = \arg\min_{\mathbf{x}} \mathbf{c}^T\mathbf{x} \tag{18.12}$$

$$\text{s.t.:} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{18.13}$$

$$\mathbf{x} \succeq 0 \tag{18.14}$$

Hàm ràng buộc cuối cùng có thể được viết lại là: $f_i(\mathbf{x}) = -x_i, i = 1, \ldots, n$. Lagrangigan của bài toán này là:

$$\mathcal{L}(\mathbf{x}, \lambda, \nu) = \mathbf{c}^T\mathbf{x} - \sum_{i=1}^{n} \lambda_i x_i + \nu^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = -\mathbf{b}^T\nu + (\mathbf{c} + \mathbf{A}^T\nu - \lambda)^T\mathbf{x}$$

(đừng quên điều kiện $\lambda \succeq 0$.) Dual function là:

$$g(\lambda, \nu) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \nu) \tag{18.15}$$

$$= -\mathbf{b}^T\nu + \inf_{\mathbf{x}}(\mathbf{c} + \mathbf{A}^T\nu - \lambda)^T\mathbf{x} \tag{18.16}$$

Nhận thấy rằng một hàm tuyến tính $\mathbf{d}^T\mathbf{x}$ của $\mathbf{x}$ bị chặn dưới khi vào chỉ khi $\mathbf{d} = 0$. Vì chỉ nếu một phần tử d_i của $\mathbf{d}$ khác 0, ta chỉ cần chọn x_i rất lớn và ngược dấu với d_i, ta sẽ có một giá trị nhỏ tuỳ ý.

Nói cách khác, $g(\lambda, \nu) = -\infty$ trừ khi $\mathbf{c} + \mathbf{A}^T\nu - \lambda = 0$. Tóm lại:

$$g(\lambda, \nu) = \begin{cases} -\mathbf{b}^T\nu & \text{if } \mathbf{c} + \mathbf{A}^T\nu - \lambda = 0 \\ -\infty & \text{otherwise} \end{cases} \tag{18.17}$$

Trường hợp thứ hai khi $g(\lambda, \nu) = -\infty$ các bạn sẽ gặp rất nhiều sau này. Trường hợp này không nhiều thú vị vì hiển nhiên $g(\lambda, \nu) \leq p^*$. Vì mục đích chính là đi tìm chặn dưới của p^* nên ta sẽ chỉ quan tâm tới các giá trị của λ và ν sao cho $g(\lambda, \nu)$ càng lớn càng tốt. Trong bài toán này, ta sẽ quan tâm tới các λ và ν sao cho $\mathbf{c} + \mathbf{A}^T\nu - \lambda = 0$.

18.4 Bài toán đối ngẫu Lagrange (The Lagrange dual problem)

Với mỗi cặp (λ, ν), hàm đối ngẫu Lagrange cho chúng ta một chặn dưới cho *optimal value* p^* của bài toán gốc (9). Câu hỏi đặt ra là: với cặp giá trị nào của (λ, ν), chúng ta sẽ có một chặn dưới tốt nhất của p^*? Nói cách khác, ta đi cần giải bài toán:

$$\lambda^*, \nu^* = \arg\max_{\lambda,\nu} g(\lambda, \nu) \tag{18.18}$$

$$\text{subject to:} \quad \lambda \succeq 0 \qquad (11) \tag{18.19}$$

Một điểm quan trọng: vì $g(\lambda, \nu)$ là *concave* và hàm ràng buộc $f_i(\lambda) = -\lambda_i$ là các hàm *convex*. Vậy bài toán (11) chính là một bài toán lồi. Vì vậy trong nhiều trường hợp, lời giải có thể dễ tìm hơn là bài toán gốc. Chú ý rằng, bài toán đối ngẫu (11) là lồi bất kể bài toán gốc (9) có là lồi hay không.

Bài toán này dược gọi là *Lagrange dual problem* (bài toán đối ngẫu Largange) ứng với bài toán (9). Bài toán (9) còn có tên gọi khác là *primal problem* (bài toán gốc). Ngoài ra, có một khái niệm nữa, gọi là *dual feasible* tức là *feasible set* của bài toán đối ngẫu, bao gồm điều kiện $\lambda \succeq 0$ và điều kiện ẩn $g(\lambda, \nu) > -\infty$ (vì ta đang đi tìm giá trị lớn nhất của hàm số nên $g(\lambda, \nu) = -\infty$ rõ ràng là không thú vị).

Nghiệm của bài toán (11), ký hiệu là λ^*, ν^* được gọi là *dual optimal* hoặc *optimal Lagrange multipliers*.

Chú ý rằng điều kiện ẩn $g(\lambda, \nu) > -\infty$, trong nhiều trường hợp, cũng có thể được viết cụ thể. Quay lại với ví dụ phía trên, điệu kiện ẩn có thể được viết thành $\mathbf{c} + \mathbf{A}^T\nu - \lambda = 0$. Đây là một hàm affine. Vì vậy, khi có thêm ràng buộc này, ta vẫn được một bài toán lồi.

18.4.1 Weak duality

Ký hiệu giá trị tối ưu của bài toán đối ngẫu (11) là d^*. Theo (11), ta đã biết rằng:

$$d^* \leq p^*$$

ngay cả khi bài toán gốc không phải là lồi.

Tính chất đơn giản này được gọi là *weak duality*. Tuy đơn giản nhưng nó cực kỳ quan trọng.

Từ đây ta quan sát thấy hai điều:

- Nếu bài toán gốc không bị chặn dưới, tức $p^* = -\infty$, ta phải có $d^* = -\infty$, tức là bài toán đối ngẫu Lagrange là *infeasible* (tức không có giá trị nào thoả mãn ràng buộc).

- Nếu bài toàn đối ngẫu là không bị chặn trên, tức $d^* = +\infty$, chúng ta phải có $p^* = +\infty$, tức bài toán gốc là *infeasible*.

Giá trị $p^* - d^*$ được gọi là *optimal duality gap* (dịch thô là *khoảng cách đối ngẫu tối ưu*). Khoảng cách này luôn luôn là một số không âm.

Đôi khi có những bài toán (lồi hoặc không) rất khó giải, nhưng ít nhất nếu ta có thể tìm được d^*, ta có thể biết được chặn dưới của bài toán gốc. Việc tìm d^* thường có thể thực hiện được vì bài toán đối ngẫu luôn luôn là lồi.

18.4.2 Strong duality và Slater's constraint qualification

Nếu đẳng thức $p^* = d^*$ thoả mãn, *the optimal duality gap* bằng không, ta nói rằng *strong duality* xảy ra. Lúc này, việc giải bài toán đối ngẫu đã giúp ta tìm được *chính xác* giá trị tối ưu của bài toán gốc.

Thật không may, *strong duality* không thường xuyên xảy ra trong các bài toán tối ưu. Tuy nhiên, nếu bài toán gốc là lồi, tức có dạng:

$$x = \arg\min_{\mathbf{x}} f_0(\mathbf{x}) \tag{18.20}$$

$$\text{subject to:} \quad f_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m \quad (12) \tag{18.21}$$

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{18.22}$$

trong đó $f_0, f_1, \ldots, f_m$ là các hàm lồi, chúng ta *thường* (không luôn luôn) có *strong duality*. Có rất nhiều nghiên cứu thiết lập các điều kiện, ngoài tính chất lồi, để *strong duality* xảy ra. Những điều kiện đó thường có tên là *constraint qualifications*.

Một trong các *constraint qualification* đơn giản nhất là *Slater's condition*.

Định nghĩa: Một điểm *feasible* của bài toán (12) được gọi là *strictly feasible* nếu:

$$f_i(\mathbf{x}) < 0, \; i = 1, 2, \ldots, m, \quad \mathbf{A}\mathbf{x} = \mathbf{b}$$

Định lý Slater: Nếu tồn tại một điểm *strictly feasible* (và bài toán gốc là lồi), thì *strong duality* xảy ra.

Điều kiện khá đơn giản sẽ giúp ích cho nhiều bài toán tối ưu sau này.

Chú ý:

- *Strong duality* không thường xuyên xảy ra. Với các bài toán lồi, việc này xảy ra thường xuyên hơn. Tồn tại những bài toán lồi mà *strong duality* không xảy ra.

- Có những bài toán không lồi nhưng *strong duality* vẫn xảy ra. Ví dụ như bài toán trong Hình 1 phía trên.

18.5 Optimality conditions

18.5.1 Complementary slackness

Giả sử rằng *strong duality* xảy ra. Gọi $\mathbf{x}^*$ là một điểm *optimal* của bài toán gốc và (λ^*, ν^*) là cặp điểm *optimal* của bài toán đối ngẫu. Ta có:

$$
\begin{aligned}
f_0(\mathbf{x}^*) &= g(\lambda^*, \nu^*) && (18.23)\\
&= \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}) + \sum_{j=1}^{p} \nu_j^* h_i(\mathbf{x}) \right) && (18.24)\\
&\leq f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) + \sum_{j=1}^{p} \nu_j^* h_j(\mathbf{x}^*) && (18.25)\\
&\leq f_0(\mathbf{x}^*) && (18.26)
\end{aligned}
$$

- Dòng đầu là do chính là *strong duality*.
- Dòng hai là do định nghĩa của hàm đối ngẫu.
- Dòng ba là hiển nhiên vì infimum của một hàm nhỏ hơn giá trị của hàm đó tại bất kỳ một điểm nào khác.
- Dòng bốn là vì các ràng buộc $f_i(\mathbf{x}^*) \leq 0, \lambda_i \geq 0, i = 1, 2, \ldots, m$ và $h_j(\mathbf{x}^*) = 0$.

Từ đây có thể thế rằng dấu đẳng thức ở dòng ba và dòng bốn phải đồng thời xảy ra. Và ta lại có thêm hai quan sát thú vị nữa:

- $\mathbf{x}^*$ chính là một điểm *optimal* của $g(\lambda^*, \nu^*)$.
- Thú vị hơn:

$$\sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) = 0$$

Vì mỗi phần tử trong tổng trên là không dương do $\lambda_i^* \geq 0, f_i \leq 0$, ta kết luận rằng:

$$\lambda_u i^* f_i(\mathbf{x}^*) = 0, i = 1, 2, \ldots, m$$

Điều kiện cuối cùng này được gọi là *complementary slackness*. Từ đây có thể suy ra:

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0 \tag{18.27}$$
$$f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0 \tag{18.28}$$

Tức ta luôn có một trong hai giá trị này bằng 0.

18.5.2 KKT optimality conditions

Chúng ta vẫn giả sử rằng các hàm đang xét có đạo hàm và bài toán tối ưu không nhất thiết là lồi.

KKT condition cho bài toán *không* lồi

Giả sử rằng *strong duality* xảy ra. Gọi $\mathbf{x}^*$ và (λ^*, ν^*) là *bất kỳ primal và dual optimal points*. Vì $\mathbf{x}^*$ tối ưu hàm khả vi $\mathcal{L}(\mathbf{x}, \lambda^*, \nu^*)$, ta có đạo hàm của Lagrangian tại $\mathbf{x}^*$ phải bằng 0.

Điều kiện Karush-Kuhn-Tucker (KKT)) nói rằng $\mathbf{x}^*, \lambda^*, \nu^*$ phải thoả mãn điều kiện:

$$f_i(\mathbf{x}^*) \leq 0, i = 1, 2, \ldots, m \tag{18.29}$$
$$h_j(\mathbf{x}^*) = 0, j = 1, 2, \ldots, p \tag{18.30}$$
$$\lambda_i^* \geq 0, i = 1, 2, \ldots, m \tag{18.31}$$
$$\lambda_i^* f_i(\mathbf{x}^*) = 0, i = 1, 2, \ldots, m \tag{18.32}$$
$$\nabla f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{j=1}^{p} \nu_j^* \nabla h_j(\mathbf{x}^*) = 0 \tag{18.33}$$

Đây là *điều kiện cần* để $\mathbf{x}^*, \lambda^*, \nu^*$ là nghiệm của hai bài toán.

KKT conditions cho bài toán lồi

Với các bài toán lồi và *strong duality* xảy ra, các điệu kiện KKT phía trên cũng là *điều kiện đủ*. Vậy với các bài toán lồi với hàm mục tiêu và hàm ràng buộc là khả vi, bất kỳ điểm nào thoả mãn các điều kiện KKT đều là *primal và dual optimal* của bài toán gốc và bài toán đối ngẫu.

Từ đây ta có thể thấy rằng: Với một bài toán lồi và điều kiện Slater thoả mãn (suy ra *strong duality*) thì các điều kiện KKT là điều cần và đủ của nghiệm.

Các điều kiện KKT rất quan trọng trong tối ưu. Trong một vài trường hợp đặc biệt (chúng ta sẽ thấy trong bài Support Vector Machine sắp tới), việc giải hệ (bất) phương trình các

điều kiện KKT là khả thi. Rất nhiều các thuật toán tối ưu được xây dựng giả trên việc giải hệ điều kiện KKT.

Ví dụ: *Equality constrained convex quadratic minimization*. Xét bài toán:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r \tag{18.34}$$

$$\text{subject to:} \quad \mathbf{A}\mathbf{x} = \mathbf{b} \tag{18.35}$$

trong đó $\mathbf{P} \in \mathbb{S}_+^n$ (tập các ma trận đối xứng nửa xác định dương).

Lagrangian:

$$\mathcal{L}(\mathbf{x}, \nu) = \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r + \nu^T(\mathbf{A}\mathbf{x} - \mathbf{b})$$

ĐIều kiện KKT cho bài toán này là:

$$\mathbf{A}\mathbf{x}^* = \mathbf{b} \tag{18.36}$$

$$\mathbf{P}\mathbf{x}^* + \mathbf{q} + \mathbf{A}^T\nu^* = 0 \tag{18.37}$$

Phương trình thứ hai chính là phương trình đạo hàm của Lagrangian tại $\mathbf{x}^*$ bằng 0.

Hệ phương trình này có thể được viết lại đơn giản là:

$$\begin{bmatrix} \mathbf{P} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \nu^* \end{bmatrix} = \begin{bmatrix} -\mathbf{q} \\ \mathbf{b} \end{bmatrix}$$

đây là một phương trình tuyến tính đơn giản!

18.6 Tóm tắt

Giả sử rằng các hàm số đều khả vi:

- Các bài toán tối ưu với chỉ ràng buộc là đẳng thức có thể được giải quyết bằng phương pháp nhân tử Lagrange. Ta cũng có định nghĩa về Lagrangian. Điều kiện cần để một điểm là nghiệm của bài toán tối ưu là nó phải làm cho đạo hàm của Lagrangian bằng 0.
- Với các bài toán tối ưu có thêm ràng buộc là bất đẳng thức (không nhất thiết là lồi), chúng ta có Lagrangian tổng quát và các biến Lagrange λ, ν. Với các giá trị (λ, ν) cố định, ta có định nghĩa về **hàm đối ngẫu Lagrange** (Lagrange dual function) $g(\lambda, \nu)$ được xác định là infimum của Lagrangian khi $\mathbf{x}$ thay đổi trên miền xác định của bài toán.
- Miền xác định và tập các điểm *feasible* thường khác nhau. *Feasible set* là tập con của tập xác định.
- Với mọi (λ, ν), $g(\lambda, \nu) \leq p^*$.

- Hàm số $g(\lambda, \nu)$ **là lồi** bất kể bài toán tối ưu có là lồi hay không. Hàm số này được gọi là *dual Lagrange fucntion* hay *hàm đối ngẫu Lagrange*.
- Bài toán đi tìm giá trị lớn nhất của hàm đối ngẫu Lagrange với điều kiện $\lambda \succeq 0$ được gọi là bài toán *đối ngẫu* (*dual problem*). Bài toán này **là lồi** bất kể bài toán gốc có lồi hay không.
- Gọi giá trị tối ưu của bài toán đối ngẫu là d^* thì ta có: $d^* \leq p^*$. Dây được gọi là *weak duality*.
- *Strong duality* xảy ra khi $d^* = p^*$. Thường thì *strong duality* không xảy ra, nhưng với các bài toán lồi thì *strong duality* thường (không luôn luôn) xảy ra.
- Nếu bài toán là lồi và điều kiện Slater thoả mãn, thì *strong duality* xảy ra.
- Nếu bài toán lồi và có *strong duality* thì nghiệm của bài toán thoả mãn các điều kiện KKT (điều kiện cần và đủ).
- Rất nhiều các bài toán tối ưu được giải quyết thông qua KKT conditions.

18.7 Kết luận

Trong ba bài 16, 17, 18, tôi đã giới thiệu *sơ lược* về tập lồi, hàm lồi, bài toán lồi, và các điệu kiện tối ưu được xây dựng thông qua *duality*. Ý định ban đầu của tôi là tránh phần này vì khá nhiều toán, tuy nhiên trong quá trình chuẩn bị cho bài Support Vector Machine, tôi nhận thấy rằng cần phải giải thích về Lagrangian - kỹ thuật được sử dụng rất nhiều trong Tối ưu. Thêm nữa, để giải thích về Lagrangian, tôi cần nói về các bài toán lồi. Chính vì vậy tôi thấy có trách nhiệm *phải* viết về ba bài này.

Trong loạt bài tiếp theo, chúng ta sẽ lại quay lại với các thuật toán Machine Learning với rất nhiều ví dụ, hình vẽ và code mẫu. Nếu bạn nào có cảm thấy hơi đuối sau ba bài tối ưu này thì cũng đừng lo, mọi chuyện rồi sẽ ổn cả thôi.

18.8 Tài liệu tham khảo

[1] Convex Optimization – Boyd and Vandenberghe, Cambridge University Press, 2004.

[2] Lagrange Multipliers - Wikipedia.