심층학습

윤예준

저자: 이안 굿펠로, 요슈아 벤지오, 에런 쿠빌

목차

3장 확률론과 정보 이론

3.11 베이즈 법칙

- $P(y|x)$를 아는 상태에서 $P(x|y)$를 구해야 할 때, $P(x)$까지만 아는 경우

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}. \tag{3.42}$$

- $P(y) = \Sigma_x P(y|x)P(x)$

3.12 연속 변수의 특별한 세부 사항

- 측도론(measure theory)
- 측도 0(measure zero) 집합
- 거의 모든 점(almost everywhere)

- 연속 변수의 또 다른 특별한 세부 사항
  - x와 y가 y = g(x) 만족한다.
  - $p_y(\mathrm{y}) = p_x(g^{-1}(y))$ 이것이 아닌, $\int p_y(y)dy = \frac{1}{2}$ (x와 y에 대해 $y = \frac{x}{2}$이고 $x \sim U(0,1)$ 이라고 하자)

$$|p_y(g(x))dy| = |p_x(x)dx|. \tag{3.44}$$

$$p_y(y) = p_x(g^{-1}(y))\left|\frac{\partial x}{\partial y}\right| \tag{3.45}$$

$$p_x(x) = p_y(g(x))\left|\frac{\partial g(x)}{\partial x}\right|. \tag{3.46}$$

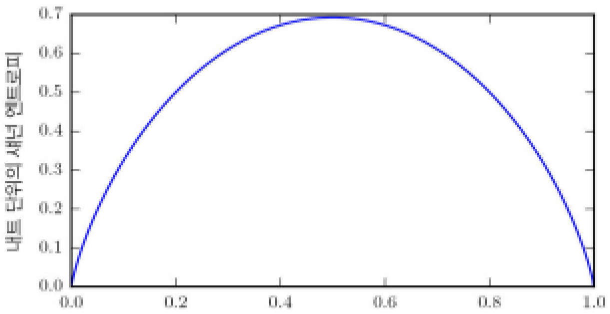
$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x}))\left|\det\left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|. \tag{3.47}$$

3.13 정보 이론

- 수량화 방법이 가져야하는 성질 3가지
  - 발생 가능성이 큰 사건은 정보량이 적어야한다. 극단적인 경우, 반드시 발생하는 사건에는 아무런 정보도 없어야 한다.
  - 발생 가능성이 낮은 사건은 정보량이 많아야 한다.
  - 개별 사건들의 정보량을 더할 수 있어야 한다.

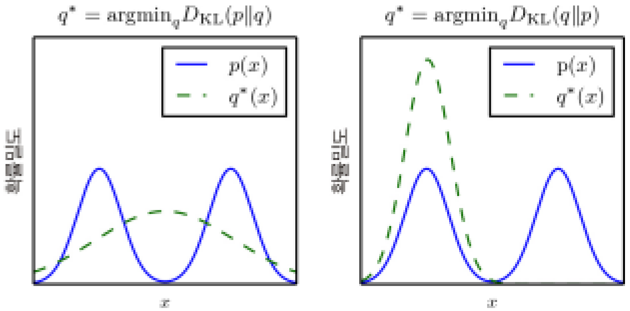
- 자기 정보: $I(x) = -\log P(x).$ (3.48)
  - 단위 nat: 1nat은 확률이 1/e인 사건을 관측해서 얻는 정보의 양

- 섀넌 엔트로피: $H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)].$ (3.49)



3.13 정보 이론

- 쿨백-라이블러 발산값(Kullback-Leibler divergence) : KL



- 교차 엔트로피

$$H(P, Q) = -\mathbb{E}_{\mathrm{x} \sim P} \log Q(x). \tag{3.51}$$

3.14 구조적 확률 모형

- 그래프: 정점들이 간선들로 연결된 구조

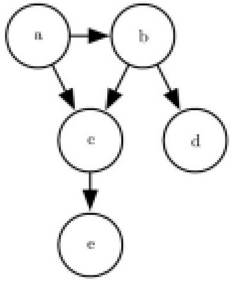
- 확률분포를 여러 인수로 분해해서 곱으로 표현 할 수 있다.

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{b}). \tag{3.52}$$

- 유향 그래프

$$p(\mathbf{x}) = \prod_i p(\mathrm{x}_i \mid Pa_{\mathcal{G}}(\mathrm{x}_i)) \tag{3.53}$$

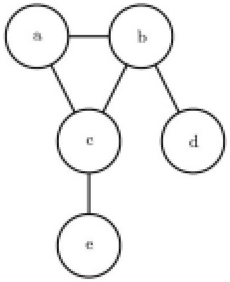
$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a}, \mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c}). \tag{3.54}$$



- 무향 그래프

$$p(\mathbf{x}) = \frac{1}{Z}\prod_i \phi^{(i)}(\mathcal{C}^{(i)}). \tag{3.55}$$

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = \frac{1}{Z}\phi^{(1)}(\mathrm{a}, \mathrm{b}, \mathrm{c})\phi^{(2)}(\mathrm{b}, \mathrm{d})\phi^{(3)}(\mathrm{c}, \mathrm{e}). \tag{3.56}$$



4장 수치 계산

4.1 Overflow and Underflow

- Uderflow: 0에 가까운 수가 반올림 때문에 정확히 0이 되는 것
- Overflow: 무한대로 근사되는 것

- 대표적인 예: 소프트맥스 함수

$$\operatorname{softmax}(\boldsymbol{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n}\exp(x_j)}. \tag{4.1}$$

4.2 Poor Conditioning

- 조건화(conditioning)
  - 입력의 작은 변화에 대해 함수가 얼마나 급하게 변하는지 뜻하는 용어

- $f(x) = A^{-1}x, A \epsilon \mathbb{R}^{n \times n}$의 조건수는 $$\max_{i,j} \left| \frac{\lambda_i}{\lambda_j} \right|. \quad (4.2)$$

4.3 Gradient-Based Optimization

- 목적함수, 판정기준
- 비용함수, 손실함수, 오차함수

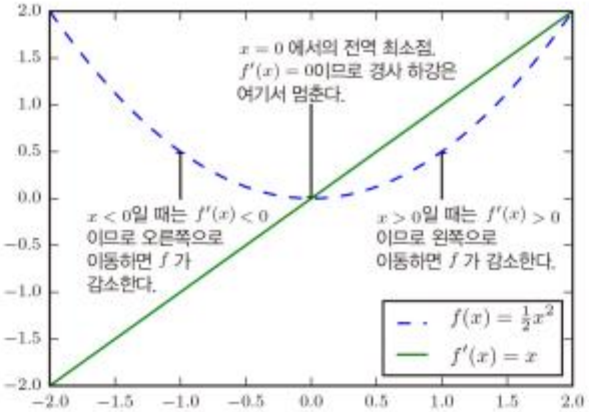


그림 4.1: 경사 하강법. 경사 하강법 알고리즘에서 함수를 따라 극소점으로 내려갈 때 이동 방향을 미분을 이용해서 판단하는 방법을 나타낸 그림이다.

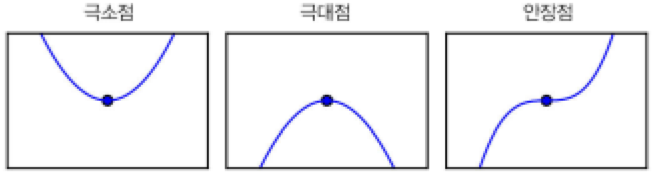


그림 4.2: 여러 종류의 임계점들. 입력이 1차원일 때의 세 가지 임계점의 예이다. 임계점은 기울기가 0인 점이다. 그런 점은 이웃 점들보다 낮은 극소점일 수도 있고, 이웃 점들보다 높은 극대점일 수도 있고, 이웃 점들보다 높기도 하고 낮기도 한 안장점일 수도 있다.

4.3 Gradient-Based Optimization

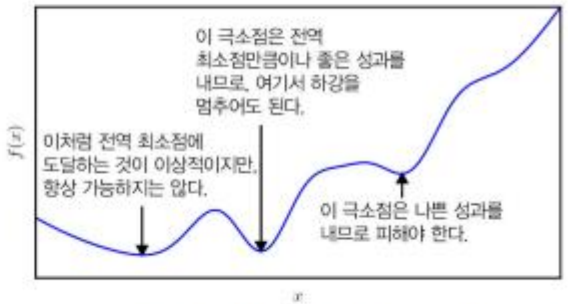


그림 4.3: 근사 최적화. 극소점이 여러 개이거나 대지(plateau; 평평한 부분)가 존재하면 최적화 알고리즘이 하나의 전역 최소점을 찾지 못할 수 있다. 심층 학습의 맥락에서는, 비용함수의 값이 현저히 낮다면 진짜 최소점이 아니라도 받아들일 때가 많다.

4.3 Gradient-Based Optimization

- 편미분 $\frac{\partial}{\partial x_i} f(x)$
  - 점 x에서 변수 x_i만 증가했을 때의 f의 변화를 측정

- 기울기 벡터
  - 미분의 개념을 입력이 여러 개인 함수로 일반화한 것
  - f의 기울기는 모든 편미분으로 구성된 벡터이고 표기 방법은 $\nabla_x f(x)$이다.
  - 다차원 입력의 임계점은 기울기 벡터의 모든 성분이 0인 점

- 방향미분
  - 함수 f의 u의 방향 기울기를 u방향의 방향미분이라 한다.

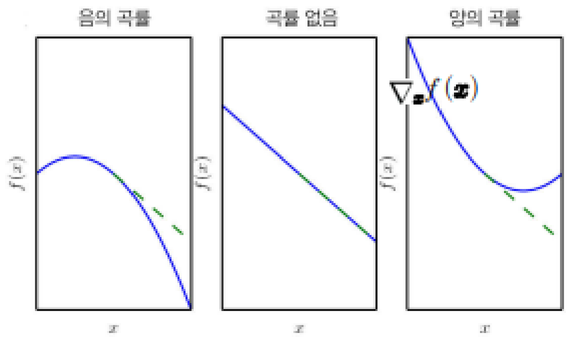
$$\min_{\boldsymbol{u}, \boldsymbol{u}^\top \boldsymbol{u} = 1} \boldsymbol{u}^\top \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \tag{4.3}$$

$$= \min_{\boldsymbol{u}, \boldsymbol{u}^\top \boldsymbol{u} = 1} \|\boldsymbol{u}\|_2 \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x})\|_2 \cos\theta. \tag{4.4}$$

$$\boldsymbol{x}' = \boldsymbol{x} - \epsilon \nabla_{\boldsymbol{x}} f(\boldsymbol{x}). \tag{4.5}$$

4.3.1 야코비 행렬과 헤세 행렬

- 야코비 행렬(Jacobian matrix)
  - 출력도 벡터인 함수의 편미분들로 모두 구해야 할 때, 그런 모든 편미분으로 구성된 행렬
  - $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$의 야코비 행렬 $J \in \mathbb{R}^{n \times m}$ 은 $J_{i,j} = \frac{\partial}{x_i} f(x)_i$로 정의된다.
- 이차미분
  - 미분의 미분
  - $f: \mathbb{R}^n \rightarrow \mathbb{R}$ 의 있을 때, f의 x_j에 대한 미분(편미분)의 x_i에 대한 미분을 $\frac{\partial^2}{\partial x_i \partial x_j} f$로 표기한다.



4.3.1 야코비 행렬과 헤세 행렬

- 헤세 행렬(Hessian matrix)
  - 함수의 입력이 다차원이면 이차미분이 여러 개이다. 그러한 이차미분들을 하나의 행렬로 모은 것
  - 헤세 행렬 $H(f)(x)$는 다음과 같이 정의 된다.

$$\boldsymbol{H}(f)(\boldsymbol{x})_{i,j} = \frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x}). \tag{4.6}$$

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\boldsymbol{x}) = \frac{\partial^2}{\partial x_j \partial x_i} f(\boldsymbol{x}). \tag{4.7}$$

- 실수 고윳값들의 집합과 고유벡터들로 이루어진 직교 기저로 분해할 수 있음 (헤세 행렬이 실숫값 대칭행렬인 경우)
  - 단위벡터 d가 가리키는 특정방향의 이차미분은 $d^T H d$로 표기
  - 최대 고윳값은 최대 이차미분 결정
  - 최소 고윳값은 최소 이차미분 결정

4.3.1 야코비 행렬과 헤세 행렬

- 현재 점 $x^{(0)}$ 주변의 함수 $f(x)$를 이차 테일러 급수로 근사할 수 있다.

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^{(0)}) + (\boldsymbol{x} - \boldsymbol{x}^{(0)})^{\top} \boldsymbol{g} + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(0)})^{\top} \boldsymbol{H}(\boldsymbol{x} - \boldsymbol{x}^{(0)}). \qquad (4.8)$$

  - g는 기울기 벡터, H는 점 $x^{(0)}$에서의 헤세 행렬

- 학습 속도가 ε이라고 할 때, 경사 하강법이 제시하는 새 점 x는 $x^{(0)} - \varepsilon g$이다.

$$f(\boldsymbol{x}^{(0)} - \epsilon \boldsymbol{g}) \approx f(\boldsymbol{x}^{(0)}) - \epsilon \boldsymbol{g}^{\top} \boldsymbol{g} + \frac{1}{2}\epsilon^2 \boldsymbol{g}^{\top} \boldsymbol{H} \boldsymbol{g}. \qquad (4.9)$$

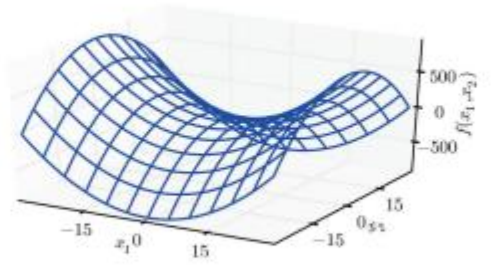
  - 함수의 원래 값
  - 함수의 기울기에 따라 예측한 개선 정도
  - 함수의 곡률에 의한 오차를 바로잡기 위한 값

- $g^T Hg$가 양수일 때, 함수의 테일러 급수 근사를 가장 많이 감소하는 최적 단계 크기

$$\epsilon^* = \frac{\boldsymbol{g}^{\top} \boldsymbol{g}}{\boldsymbol{g}^{\top} \boldsymbol{H} \boldsymbol{g}}. \qquad (4.10)$$

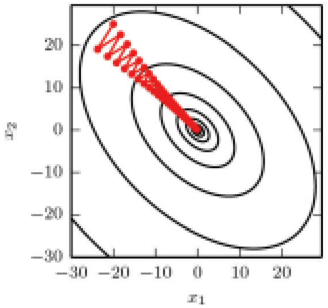
4.3.1 야코비 행렬과 헤세 행렬

- *이차미분 판정*
  - $f'(x) = 0$이고 $f''(x) > 0$일 때는 x가 극소점
  - $f'(x) = 0$이고 $f''(x) < 0$일 때는 x가 극대점
  - $f''(x) = 0$일 때는 판정의 결론이 나지 않는다.
    - x는 안장점이나 평평한 지역의 일부일 가능성 존재



4.3.1 야코비 행렬과 헤세 행렬

- 헤세 행렬의 조건수가 나쁜 경우
  - 성과가 좋지 않음.
  - 한 방향에서는 미분이 빠르게 증가하지만 다른 방향에서는 느리게 증가하기 때문
  - 경사하강법은 이러한 미분들의 변화를 알아채지 못함



- 뉴턴법
  - 위의 문제를 해결한 방법
  - 이차 테일러 급수 전개를 이용하여 어떤 점 $x^{(0)}$ 근처의 $f(x)$를 근사한다.

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^{(0)}) + (\boldsymbol{x} - \boldsymbol{x}^{(0)})^\top \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(0)}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(0)})^\top \boldsymbol{H}(f)(\boldsymbol{x}^{(0)})(\boldsymbol{x} - \boldsymbol{x}^{(0)}). \quad (4.11)$$

$$\boldsymbol{x}^* = \boldsymbol{x}^{(0)} - \boldsymbol{H}(f)(\boldsymbol{x}^{(0)})^{-1} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(0)}). \quad (4.12)$$

4.3.1 야코비 행렬과 헤세 행렬

- 일차 최적화 알고리즘
- 이차 최적화 알고리즘
- 립시츠 연속
  - 립시츠 연속 함수란 변화율의 범위가 립시츠 상수$\mathcal{L}$로 결정되는 함수 f를 말한다.

$$\forall \boldsymbol{x}, \forall \boldsymbol{y}, |f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \mathcal{L} \|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{4.13}$$

- 볼록함수 최적화

4.4 Constrained Optimization

- 함수 $f(x)$를 x의 모든 가능한 값에 대해 최대하 또는 최소화하는 것이 아니라, 어떤 집합 $\mathbb{S}$에 속한 x의 값들에 대해서만 $f(x)$를 최대하 또는 최소화하고 싶을 때를 말한다.
- 집합 $\mathbb{S}$에 속하는 점 x들을 제약 최적화의 용어로 **실현 가능 점**이라 한다.

- 제약 접근 방식
  - 작은 해를 찾아야하는 경우, 가장 흔히 쓰이는 접근 방식 $||x|| \leq 1$ 같은 크기 제약을 두는 것
  - 간단한 제약 최적화 접근 방식 중 하나는, 그냥 주어진 제약을 고려하도록 경사 하강법 알고리즘을 수정하는 것
  - 더 정교한 접근 방식은 그 해(최종 결과)를 원래의 제약 최적화 문제의 해로 변환 할 수 있는 또 다른 무제약 최적화 문제를 고안하는 것

- 캐러시-쿤-터커(Karush-Kuhn-Tucker) 접근 방식 (KKT)
  - 제약 최적화에 대한 아주 일반적인 해법 제공
  - KKT접근 방식에서는 **일반화된 라그랑주 함수**라고 하는 새로운 종류의 함수를 사용

4.4 Constrained Optimization

- $\mathbb{S} = \{x | \forall\, i, g^{(i)}(x) = 0$ 그리고 $\forall\, j, h^{(j)}(x) \le 0\}$
  - $g^{(i)}$가 관여하는 등식을 **상등 제약**
  - $h^{(j)}$가 관여하는 부등식을 **부등 제약**
- KTT승수 요소들로 서술된 일반화된 라그랑주 함수 정의

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) + \sum_i \lambda_i g^{(i)}(\boldsymbol{x}) + \sum_j \alpha_j h^{(j)}(\boldsymbol{x}). \tag{4.14}$$

- 최소화 문제 해결 방안

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \ge 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \tag{4.15}$$

$$\min_{\boldsymbol{x} \in \mathbb{S}} f(\boldsymbol{x}). \tag{4.16}$$

$$\max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \ge 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) \tag{4.17}$$

$$\max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \ge 0} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = \infty \tag{4.18}$$

4.4 Constrained Optimization

- 최대화가 목표인 제약 최적화

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq 0} -f(\boldsymbol{x}) + \sum_i \lambda_i g^{(i)}(\boldsymbol{x}) + \sum_j \alpha_j h^{(j)}(\boldsymbol{x}). \tag{4.19}$$

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{\lambda}} \max_{\boldsymbol{\alpha}, \boldsymbol{\alpha} \geq 0} f(\boldsymbol{x}) + \sum_i \lambda_i g^{(i)}(\boldsymbol{x}) - \sum_j \alpha_j h^{(j)}(\boldsymbol{x}). \tag{4.20}$$

- 활성 (active)
  - $h^{(i)}(x^*) = 0$인 제약 $h^{(i)}(x)$
- $\alpha \odot h(x) = 0$성립
  - 비활성 $h^{(i)}$의 값은 음수이므로, $min_x min_\lambda max_{\alpha,\alpha \geq 0} L(x, \lambda, \alpha)$의 해에서 $\alpha_i = 0$이다.

4.4 Constrained Optimization

- 캐러시-쿤-터커(KKT) 조건
  - 제약 최적화 문제의 최적점들을 많지 않은 수의 성질들로 서술 할 수 있는 성질들의 집합
    - 일반화된 라그랑주 함수의 기울기가 0이다
    - x와 KTT 승수에 대한 모든 제약을 충족한다.
    - 부등 제약은 '상보적 여유 조건 ' 을 나타낸다. 즉 $\alpha \odot h(x) = 0$이다.

4.5 Example: Linzer Least Squares

- 식 4.21의 함수를 최소화하는 x의 값을 구한다고 하자.

$$f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2. \tag{4.21}$$

  - 경사 하강법 기반 최적화 기법

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top\boldsymbol{b}. \tag{4.22}$$

  - 뉴턴법

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda(\boldsymbol{x}^\top\boldsymbol{x} - 1). \tag{4.23}$$

$$\min_{\boldsymbol{x}} \max_{\lambda, \lambda \geq 0} L(\boldsymbol{x}, \lambda). \tag{4.24}$$

4.5 Example: Linzer Least Squares

- 무어-펜로즈 유사 역행렬

$$\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top \boldsymbol{b} + 2\lambda \boldsymbol{x} = 0. \tag{4.25}$$

$$\boldsymbol{x} = (\boldsymbol{A}^\top \boldsymbol{A} + 2\lambda \boldsymbol{I})^{-1} \boldsymbol{A}^\top \boldsymbol{b}. \tag{4.26}$$

$$\frac{\partial}{\partial \lambda} L(\boldsymbol{x}, \lambda) = \boldsymbol{x}^\top \boldsymbol{x} - 1. \tag{4.27}$$