creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이학석사 학위논문

이차 적률 행렬이 갱신되는 변수에 의존하는 경우의 적률 조정 근위 경사하강법 수렴 분석

2022년 02월

서울대학교 대학원

통계학과

안 우 진

이차 적률 행렬이 갱신되는 변수에 의존하는 경우의 적률 조정 근위 경사하강법 수렴 분석

지도교수 원 중 호

이 논문을 이학석사 학위논문으로 제출함
2021년 10월

서울대학교 대학원
통계학과
안 우 진

안우진의 석사 학위논문을 인준함
2022년 01월

위 원 장 _____ 장 원 철 _____ (인)

부위원장 _____ 원 중 호 _____ (인)

위 원 _____ 이 권 상 _____ (인)

국문초록

본 논문에서는 매끄럽지 않은 정규화 항이 있는 목적함수를 최적화하는 MadProx 방법을 소개하고 그 수렴성을 분석한다. MadProx에 앞서, MadProx의 근간이 되는 방법인 MasGrad 방법을 소개하고 MasGrad의 통계적 추론 성질을 증명할 것이다. MadProx 방법은 이차 적률 행렬이 갱신되는 변수에 의존하지 않는 경우와 같은 제한된 조건에서 수렴한다는 한계가 있었다. 본 논문에서는 새로운 검색 방향을 정의하고 고정된 단계-크기가 아닌 직선 탐색 과정을 통해 매 반복마다 단계-크기를 정하는 알고리즘을 소개할 것이다. 그렇게 구한 검색 방향과 단계-크기로 이차 적률 행렬이 갱신되는 변수에 의존하는 경우에도 MadProx가 수렴함을 보일 것이다. 더불어, 데이터 시뮬레이션 환경에서 L1-노름 정규화 항이 있는 로지스틱 회귀를 대표 모형으로 하여 MadProx의 수렴 속도를 다른 반복법들과 비교할 것이다.

주요어: 최적화, 반복법, MasGrad, MadProx, 근위 함수, 직선 탐색 과정
학 번: 2020-21707

Contents

1 서론 **2**

1.1 반복법 2

1.2 기여 2

2 MasGrad의 통계적 추론 **4**

2.1 Langevin Diffusion 5

2.2 비점근적 상계 5

3 MasGrad의 수렴 분석 **9**

3.1 Non-convex case 9

3.2 Strongly convex case 11

3.3 일반화 선형 모형에서의 가속 15

4 매끄럽지 않은 정규화 문제 **19**

4.1 ProxNewton 20

4.2 MadProx 20

5 MadProx의 성질 **21**

5.1 검색 방향 21

5.2 직선 탐색 과정 24

6 MadProx의 수렴성 **28**

6.1 $\mathbf{V}(x_t) = \mathbf{V}$일 때 MadProx의 수렴성 28

6.2 $\mathbf{V}(x_t)$가 x_t에 의존할 때 MadProx의 전역 수렴성 31

6.3 $\mathbf{V}(x_t)$가 x_t에 의존할 때 지역 수렴성 32

7 실험 **34**

7.1 실험 환경 34

7.2 실험 결과 35

8 결론 **39**

9 부록 **40**

1 서론

1.1 반복법

많은 통계적 방법은 모수 θ에 대한 목적함수 $L(\theta)$를 최적화하는 문제들로 구성된다. 이 문제를 최적화 식으로 나타내면 다음과 같다:

$$\underset{\theta \in \mathbb{R}^n}{\text{minimize}} \;\; L(\theta). \tag{1}$$

예를 들어, 로지스틱 회귀 모형의 경우 자료 $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, $(i = 1, \cdots, N)$로부터 구한 목적함수 $L(\theta) = \prod_{i=1}^{N} f(y_i|\mathbf{x}_i, \theta)$를 최대화하는 모수 θ를 구하는 문제로 표현할 수 있다. 목적함수 (1)을 최적화하는 대표적인 방법으로 선형 계획법이나 최소제곱법과 같이 직접적으로 목적함수의 최적값을 계산하는 방법과 반복적으로 최적점의 근삿값을 갱신하는 방법인 반복법이 있다. 본 논문에서는 아래와 같은 반복법을 중점적으로 다룰 것이다:

$$\theta_{t+1} = \theta_t - \eta_t \mathbf{h}(\theta_t), \tag{2}$$

여기서 $\mathbf{h}$는 목적함수의 일차 혹은 이차 정보에 의존하는 함수이고 η_t는 얼마나 갱신할지에 대한 단계-크기(step-size)이다.

1.2 기여

반복법에는 목적함수의 일차 정보만을 이용하는 방법과 이차 정보를 추가로 이용하는 방법이 있다. 일차 정보만을 이용하는 방법에는 경사하강법(gradient descent method, 이하 GD)과 확률적 경사하강법(stochastic gradient descent method, 이하 SGD) 등이 있고 이차 정보를 이용하는 방법에는 뉴턴 방법(Newton's method)이 널리 알려져 있다. 본 논문에서는 이차 정보를 이용하는 방법 중 헤시안 행렬을 사용하는 뉴턴 방법이 아닌 이차 적률 행렬을 사용하는 적률 조정 확률적 경사하강법(moment adjusted stochastic gradient descent method, 이하 MasGrad)이라는 방법에 대해서 소개하고 해당 방법의 특징들을 알아 볼 것이다.

더 나아가, 제약 조건이 있는 문제를 이차 적률 행렬을 사용하여 최적화하는 방법인 적률 조정 근위 경사하강법(moment adjusted proximal gradient descent method, 이하 MadProx)을 소개하고 그 수렴성을 분석할 것이다. MadProx를 소개한 [4]에는 이차 적률 행렬이 갱신되는 변수에 의존하지 않는 제한된 조건에서 수렴 분석을 진행한 반면, 본 논문에는 이차 적률 행렬이 갱신되는 변수에 의존하는 경우에서 수렴 분석을 할 것이다. 그리고 매 반복마다 직선 탐색 과정을 통해 단계-크기 η_t를 구하면 MadProx가 1차 수렴 속도보다 더 빠른 2차 수렴 속도를 가지는 것을 실험을 통해 보일 것이다. 즉, 여러 데이터 실험 환경에서 $\ell 1$-노름 정규화 항이 있는 로지스틱 회귀 모형을 대표 모형으로 하여, 근위 뉴턴 방법(proximal-newton type method, 이하 ProxNewton), 좌표 하강법(coordinate descent method, 이하 CD)과 MadProx의 수렴성을 비교할 것이다.

본 논문은 주로 참고문헌 [3]과 [4]를 참조하였다. 그 중 참고문헌 [4]의 정리들을 이해하기

쉽게 다시 작성하였고 일부 증명 과정에서는 표기에 오류가 있어서 수정하여 다시 증명하였다. 또한, 참고문헌 [3]에서의 증명을 MadProx 버전으로 수정하였다.

2 MasGrad의 통계적 추론

목적함수가 $L(\theta) = \mathbb{E}_{\mathbf{z}\sim P}\left[\ell(\theta, \mathbf{z})\right]$일 때, (2)에서 $\mathbf{h}(\theta_t) = \mathbb{E}_{\mathbf{z}\sim P}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]$인 경우를 경사하강법(gradient descent method)이라고 한다. 즉, 경사하강법은 매 반복마다 현재 값에 목적함수의 기울기 값을 빼어서 다음 단계의 값으로 갱신하는 반복법이다. 경사하강법은 목적함수의 극값을 찾는 제일 간단하고 기본적인 일차 반복법이다.

$\mathbb{E}_{\mathbf{z}\sim P}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]$를 직접 계산하는 대신 데이터의 모분포 $z \sim P$에서 독립적으로 추출한 n개의 표본을 이용해 $\hat{\mathbb{E}}_n\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]$를 계산하면 경사하강법과 마찬가지로 그 수렴값이 극값에 수렴하고 반복마다 기울기의 계산량이 줄어든다는 점에서 유용하다. 이 방법을 확률적 경사 하강법(stochastic gradient descent method)이라고 한다:

$$\theta_{t+1} = \theta_t - \eta_t \hat{\mathbb{E}}_n\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right],$$

여기서 $\hat{\mathbb{E}}_n\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(\theta_t, \mathbf{z}_i)$은 독립적으로 추출된 n개의 표본에 대한 경험적 기댓값이다.

한편, (2)에서 $\mathbf{h}(\theta_t) = \mathbf{V}(\theta_t)^{-1}\hat{\mathbb{E}}_n\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]$인 경우를 적률 조정 확률적 경사하강법(moment adjusted stochastic gradient descent method, MasGrad [4])이라고 한다:

$$\theta_{t+1} = \theta_t - \eta_t \mathbf{V}(\theta_t)^{-1}\hat{\mathbb{E}}_n\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right], \tag{3}$$

여기서 $\mathbf{V}(\theta_t) = \left\{\mathrm{Cov}_{\mathbf{z}\sim P}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]\right\}^{1/2}$는 점수 함수(score function) $\nabla_\theta \ell(\theta_t, \mathbf{z})$에 대한 공분산 행렬의 제곱근이다. 위 방법은 SGD에서 $\hat{\mathbb{E}}_n$을 표준화한 방법으로서 각 좌표의 조건부 노이즈가 독립적이고 동질적(homogeneous)이게 된다. 이러한 표준화 기법은 더 노이즈가 많은 성분이 더 많은 조정을 받는다는 점에서 다른 이차 반복법인 뉴턴 방법과 유사하다.

2장에서는 MasGrad로부터 유도된 확률적 갱신 ξ_t(아래의 diff_MasGrad)의 분포가 적절한 조건하에서 MasGrad 갱신 θ_t의 분포와 가까운 거리에 있음을 보일 것이다. 다시 말해서, 데이터의 모분포 $\mathbf{z} \sim P$와 관계없이 두 갱신의 분포 사이의 거리는 비점근적 상계를 가진다.

$$\text{MasGrad}: \theta_{t+1} = \theta_t - \eta \mathbf{V}(\theta_t)^{-1}\hat{\mathbb{E}}_n\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] \tag{4}$$

$$\text{diff_MasGrad}: \xi_{t+1} = \xi_t - \eta \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\,\mathbf{g}_t \tag{5}$$

여기서 $\mathbf{b}(\theta) = \mathbb{E}_{\mathbf{z}\sim P}\left[\nabla_\theta \ell(\theta, \mathbf{z})\right]$, $\mathbf{V}(\theta) = \left\{\mathrm{Cov}\left[\nabla_\theta \ell(\theta, \mathbf{z})\right]\right\}^{1/2}$, $\beta = \frac{2n}{\eta}$ 그리고 $\mathbf{g}_t \sim \mathcal{N}(0, I_p)$이다.

이 장의 내용은 [4]의 결과를 알기 쉽게 정리한 것이다.

2.1 Langevin Diffusion

MasGrad로부터 유도된 확률적 갱신 (5)를 discretized Langevin diffusion 혹은 diff_MasGrad 라고 한다. MasGrad에서 중심극한정리를 통해 다음과 같은 확률 변수가 있는 근사식을 유도 할 수 있다:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta \mathbf{V}(\theta_t)^{-1} \hat{\mathbb{E}}_n \left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] \\ &= \theta_t - \eta \mathbf{V}(\theta_t)^{-1} \mathbb{E}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] + \eta \mathbf{V}(\theta_t)^{-1} \left[\mathbb{E}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] - \hat{\mathbb{E}}_n \left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]\right] \\ &\approx \theta_t - \eta \mathbf{V}(\theta_t)^{-1} \mathbf{b}(\theta_t) + \sqrt{2\beta^{-1}\eta}\, \mathbf{g}_t,\end{aligned}$$

여기서 $\mathbf{b}(\theta) = \mathbb{E}_{\mathbf{z}\sim P}\left[\nabla_\theta \ell(\theta, \mathbf{z})\right]$, $\mathbf{V}(\theta) = \left\{\mathrm{Cov}\left[\nabla_\theta \ell(\theta, \mathbf{z})\right]\right\}^{1/2}$, $\beta = \frac{2n}{\eta}$ 그리고 $\mathbf{g}_t \sim \mathcal{N}(0, I_p)$ 이다.

Proposition 1.

$$\eta \mathbf{V}(\theta_t)^{-1} \left(\mathbb{E}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] - \hat{\mathbb{E}}_n \left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]\right) \approx \sqrt{2\beta^{-1}\eta}\, \mathbf{g}_t$$

Proof. 아래와 같이 $X_i(\theta_t)$를 정의하면 $\mathbb{E}\left[X_i(\theta_t)\right] = 0$, $\mathrm{Cov}\left[X_i(\theta_t)\right] = I_p$이 성립한다:

$$i = 1, 2, 3, \cdots, n \text{ 에 대해 } X_i(\theta_t) = \mathbf{V}(\theta_t)^{-1} \left(\nabla_\theta \ell(\theta_t, z_i) - \mathbb{E}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]\right).$$

여기서 $X_i(\theta_t)$는 분포 P에서 추출된 z_i와 같은 σ-field에서 정의된다. 중심극한정리에 의해,

$$\begin{aligned}&\sqrt{n}\mathbf{V}(\theta_t)^{-1} \left(\hat{\mathbb{E}}_n \left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] - \mathbb{E}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]\right) \\ &= \sqrt{n} \left(\frac{1}{n}\sum_{i=1}^{n} X_i(\theta_t) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i(\theta_t)\right]\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_p).\end{aligned}$$

따라서, $\eta \mathbf{V}(\theta_t)^{-1}(\mathbb{E}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] - \hat{\mathbb{E}}_n \left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]) \approx \sqrt{2\beta^{-1}\eta}\, \mathbf{g}_t$, 여기서 $\beta = \frac{2n}{\eta}$, $\mathbf{g}_t \sim \mathcal{N}(0, I_p)$ 이다. □

2.2 비점근적 상계

MasGrad 갱신 (4)으로부터 얻어진 $\mu(\theta_t, t \in [T])$를 MasGrad 확률 과정에 대한 결합 분포, diff_MasGrad 갱신 (5)로부터 얻어진 $\mu(\xi_t, t \in [T])$를 diff_MasGrad 확률 과정에 대한

결합 분포라고 하자. 적절한 가정하에 $\mu(\theta_t, t \in [T])$와 $\mu(\xi_t, t \in [T])$의 총 변동 거리가 비점근적 상계를 가진다. 다시 말해서, 충분히 큰 n에 대하여 $\mu(\xi_t, t \in [T])$는 $\mu(\xi_t, t \in [T])$에 분포 수렴한다. 즉,

$$D_{\mathrm{TV}}(\mu(\theta_t, t \in [T]),\ \mu(\xi_t, t \in [T])) \leq O_{t,n}\left(\sqrt{\frac{t}{n}}\right)$$
$$\Rightarrow \quad \mu(\theta_t, t \in [T]) \xrightarrow{\mathcal{L}} \mu(\xi_t, t \in [T]) \text{ as } n \to \infty.$$

Lemma 1. ***(Pinkser's inequality)*** *([5]의 Lemma 2.5)*

임의의 확률 변수 X, Y에 대해

$$\frac{1}{2} D_{TV}(\mu(X), \mu(Y))^2 \leq D_{KL}(\mu(X) \| \mu(Y)). \tag{6}$$

Theorem 1. ***(Non-asymptotic bound for inference)*** *([4]의 Theorem 3.1)*

$i = 1, 2, 3, \cdots, n$에 대해 $X_i(\theta_t) = \mathbf{V}(\theta_t)^{-1}\left[\nabla_\theta \ell(\theta_t, z_i) - \mathbb{E}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]\right]$이라고 하자. 임의의 θ_t에 대해 $X_i(\theta_t)$가 부록의 (1)과 (2)를 만족한다고 가정하자. MasGrad 갱신 θ_t와 diff_MasGrad 갱신 ξ_t의 초깃값이 같을 때, 즉 $\theta_0 = \xi_0$, 다음의 상계가 성립한다.

$$D_{TV}(\mu(\theta_t, t \in [T]),\ \mu(\xi_t, t \in [T])) \leq C\sqrt{\frac{T}{n} + o\left(\frac{T(\log n)^{\frac{p-4+\delta}{2}}}{n^{1+\frac{\delta}{2}}}\right)},$$

여기서 C는 D_1와 D_2에만 의존하는 상수다.

Proof.

$$\begin{aligned}
\theta_{t+1} &= \theta_t - \eta \mathbf{V}(\theta_t)^{-1} \hat{\mathbb{E}}_n\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] \\
&= \theta_t - \eta \mathbf{V}(\theta_t)^{-1} \mathbb{E}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] + \eta \mathbf{V}(\theta_t)^{-1}\left(\mathbb{E}\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right] - \hat{\mathbb{E}}_n\left[\nabla_\theta \ell(\theta_t, \mathbf{z})\right]\right) \\
&= \theta_t - \eta \mathbf{V}(\theta_t)^{-1} \mathbf{b}(\theta_t) + \sqrt{2\beta^{-1}\eta}\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(\theta_t)
\end{aligned}$$

$S_n(X, \theta_t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(\theta_t)$ 이라고 하자.

부록의 가정 (1)과 (2) 하에, [1]의 Theorem 6.1에 의하면 각 단계 t마다 다음이 성립한다.

어떤 상수 $C > 0$ 에 대해

$$D_{\mathrm{KL}}(\mu(S_n(X, \theta_t)) \| \mu(\mathbf{g_t}) | \theta_t) = \frac{C}{n} + o\left(\frac{(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{1+\frac{\delta}{2}}}\right).$$

두 확률 과정의 relative entropy에 대해 연쇄 법칙을 적용하면 다음과 같다.

$$
\begin{aligned}
&D_{\mathrm{KL}}(\mu(\theta_t, t\in[T])\|\mu(\xi_t, t\in[T]))\\
&\overset{(*)}{=} D_{\mathrm{KL}}(\mu(\theta_t, t\in[T-1])\|\mu(\xi_t, t\in[T-1]))\\
&\quad+\int D_{\mathrm{KL}}(\mu(S_n(X,\theta_{T-1}))\|\mu(\mathbf{g}_{T-1})|\theta_{T-1})\mathrm{d}\mu(\theta_t, t\in[T-1])\\
&\le D_{\mathrm{KL}}(\mu(\theta_t, t\in[T-1])\|\mu(\xi_t, t\in[T-1]))+o\left(\frac{(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{1+\frac{\delta}{2}}}\right)\\
&\le D_{\mathrm{KL}}(\mu(\theta_t, t\in[T-2])\|\mu(\xi_t, t\in[T-2]))\\
&\quad+\int D_{\mathrm{KL}}(\mu(S_n(X,\theta_{T-2}))\|\mu(\mathbf{g}_{T-2})|\theta_{T-2})\mathrm{d}\mu(\theta_t, t\in[T-2])+o\left(\frac{(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{1+\frac{\delta}{2}}}\right)\\
&\le\cdots\le D_{\mathrm{KL}}(\mu(\theta_0)\|\mu(\xi_0))+\frac{CT}{n}+o\left(\frac{T(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{1+\frac{\delta}{2}}}\right)=\underbrace{\frac{CT}{n}+o\left(\frac{T(\log n)^{\frac{p-(4+\delta)}{2}}}{n^{1+\frac{\delta}{2}}}\right)}_{(**)}.
\end{aligned}
$$

$(*)$에는 relative entropy에 대한 연쇄 법칙과 다음의 사실을 이용했다.

$$a, b<0 \text{ 에 대해, } D_{\mathrm{KL}}(\mu(a+bX)\|\mu(a+bY))=D_{\mathrm{KL}}(\mu(X)\|\mu(Y)).$$

즉,

$$
\begin{aligned}
&D_{\mathrm{KL}}(\mu(\theta_T)\|\mu(\xi_T))\\
&=D_{\mathrm{KL}}\left(\mu\left(\theta_{T-1}-\eta\mathbf{V}(\theta_{T-1})^{-1}\mathbf{b}(\theta_{T-1})+\sqrt{2\beta^{-1}\eta}\,\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i(\theta_{T-1})\right)\middle\|\right.\\
&\left.\mu\left(\theta_{T-1}-\eta\mathbf{V}(\theta_{T-1})^{-1}\mathbf{b}(\theta_{T-1})+\sqrt{2\beta^{-1}\eta}\,\mathbf{g}_{T-1}\right)|\theta_{T-1}\right)\\
&=D_{\mathrm{KL}}\left(\mu\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i(\theta_{T-1})\right)\middle\|\mu(\mathbf{g}_{T-1})\middle|\theta_{T-1}\right)\\
&=D_{\mathrm{KL}}(\mu\left(S_n(X,\theta_{T-1})\right)\|\mu(\mathbf{g}_{T-1})|\theta_{T-1})\,.
\end{aligned}
$$

$(**)$에 (6)을 적용하면 두 확률 과정의 총 변동 거리의 상계을 구할 수 있다.
즉,

$$D_{\mathrm{TV}}(\mu(\theta_t, t\in[T]),\ \mu(\xi_t, t\in[T]))\le C\sqrt{\frac{T}{n}+o\left(\frac{T(\log n)^{\frac{p-4+\delta}{2}}}{n^{1+\frac{\delta}{2}}}\right)}.$$

□

Remark. 위에서 구한 비점근적 상계의 결과에 의해, MasGrad 갱신 θ_t의 결합 분포와 diff_MasGrad

갱신 ξ_t의 결합 분포가 가깝다는 것을 알 수 있다. 따라서, MasGrad 갱신 θ_t의 분포의 특성을 확률적 갱신인 diff_MasGrad 갱신 ξ_t의 분포를 이용하여 추론할 수 있다.

3 MasGrad의 수렴 분석

이 장에서는 목적함수 $L(\theta)$가 강볼록(strongly convex) 함수인 경우와 볼록이 아닌(non-convex) 경우에 MasGrad 갱신 θ_t의 수렴성을 분석할 것이다. 특히, 일반화 선형 모형(General Linear Models, GLMs)의 경우 적절한 조건하에 MasGrad의 표준화 아이디어는 네스테로프 가속(Nesterov acceleration) 성질을 가진다. 즉, 경사하강법과 MasGrad의 경우 ϵ-minimizer에 대한 총 반복 횟수 T는 다음과 같다:

$$T_{\mathrm{GD}} = O_{\epsilon,\kappa}\left(\kappa \log \frac{1}{\epsilon}\right),$$
$$T_{\mathrm{MasGrad}} = O_{\epsilon,\kappa}\left(\sqrt{\kappa} \log \frac{1}{\epsilon}\right), \text{ for some } \kappa > 1.$$

이 장의 내용 역시 [4]의 결과를 알기 쉽게 정리한 것이다.

3.1 Non-convex case

$L(\theta) = \mathbb{E}_{\mathbf{z}\sim P}\ell(\theta, \mathbf{z})$는 매끄러운 비볼록 함수(smooth non-convex function), $\mathbf{b}(\theta) = \nabla L(\theta)$, $\mathbf{H}(\theta) = \nabla^2 L(\theta)$ 그리고 $\mathbf{V}(\theta) = \{\mathrm{Cov}\left[\nabla_\theta \ell(\theta, \mathbf{z})\right]\}^{1/2} \in \mathbb{R}^{p\times p}$가 양의 정부호 행렬이라고 하자. 그리고 아래와 같이 γ를 정의하자.

$$\gamma := \max_{v,w}\ \lambda_{\max}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right) > 0.$$

이때 다음의 정리가 성립한다.

Theorem 2. ***(MasGrad: non-convex)*** *([4]의 Theorem 5.1)*

γ는 위에서 정의한 것이라고 하자. 그리고 단계-크기가 $\eta = 1/\gamma$라고 할 때, MasGrad 갱신 θ_t와 diff_MasGrad ξ_t를 고려하자. 그러면 임의의 정밀도(precision) $\epsilon, \delta > 0$에 대해

$$(1)\ D_{TV}(\mu(\theta_t, t \in [T]),\ \mu(\xi_t, t \in [T])) \le O_\delta(\delta),$$
$$(2)\ \mathbb{E}\min_{t\le T}\ \|\nabla L(\theta_t)\| \le \epsilon,\ \mathbb{E}\min_{t\le T}\ \|\nabla L(\xi_t)\| \le \epsilon$$

를 만족하는 T, n은 다음과 같다:

$$T = \frac{2\gamma(L(\xi_0) - \min_\xi\ L(\xi)) + p\delta^2}{\epsilon^2} \cdot \left(\max_\xi \|\mathbf{V}(\xi)\| \vee 1\right),\ n = \frac{T}{\delta^2}.$$

그리고 독립적인 표본의 개수는 총 $O_{\epsilon,\delta}(\epsilon^{-4}\delta^{-2})$이다.

Proof. 편의상 $\mathbf{H}_{\tilde{c}} = \mathbf{H}(\tilde{c}\xi_t + (1-\tilde{c})\xi_{t+1})$, $C = \max_\xi \ \|\mathbf{V}(\xi)\|$라고 하자. 테일러 정리에 의해, $\tilde{c} \in [0,1]$에 대해 다음 식이 성립한다.

$$
\begin{aligned}
&\mathbb{E}[L(\xi_{t+1})|\xi_t] \\
&\overset{(*)}{=} \mathbb{E}\left[L(\xi_t) + \langle \mathbf{b}(\xi_t), \xi_{t+1} - \xi_t\rangle + \frac{1}{2}(\xi_{t+1}-\xi_t)^T \mathbf{H}(\tilde{c}\xi_t + (1-\tilde{c})\xi_{t+1})(\xi_{t+1}-\xi_t)\middle|\xi_t\right] \\
&\overset{(**)}{\leq} L(\xi_t) - \eta\left\langle \mathbf{b}(\xi_t), \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t)\right\rangle + \mathbb{E}\left[\frac{\gamma}{2}\left\|\eta\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t) - \sqrt{2\beta^{-1}\eta}\mathbf{V}(\xi_t)^{\frac{1}{2}}\mathbf{g}_t\right\|^2\middle|\xi_t\right] \\
&\leq L(\xi_t) - \eta\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2 + \frac{\eta^2\gamma}{2}\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2 + \beta^{-1}\eta\gamma\mathbb{E}[\left\|\mathbf{V}(\xi_t)^{\frac{1}{2}}\mathbf{g}_t\right\|^2|\xi_t] \\
&\leq L(\xi_t) - \frac{1}{2\gamma}\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2 + C^{\frac{1}{2}}\cdot p\beta^{-1}\eta\gamma. \qquad (7)
\end{aligned}
$$

위 부등식의 (*)와 (**)는 아래의 식을 이용했다.

$(*)$ 테일러 정리

$(**)$ $\xi_{t+1} - \xi_t = -\eta\mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\ \mathbf{g}_t$, $\gamma = \max_{v,w} \lambda_{\max}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right)$

(7)의 양변에 ξ_t에 대한 기댓값을 취하고 $t=0$부터 $T-1$까지 차례로 더하면 다음 식이 나온다.

$$
\begin{aligned}
L(\xi_0) - \min_\xi \ L(\xi) + C^{\frac{1}{2}}\cdot p\beta^{-1}\eta\gamma T &\geq \sum_{t=0}^{T-1}\frac{1}{2\gamma}\mathbb{E}\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2, \\
\mathbb{E}\min_{t\leq T}\ \left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2 &\leq \frac{L(\xi_0) - \min_\xi \ L(\xi) + C^{\frac{1}{2}}\cdot p\beta^{-1}\eta\gamma T}{T/2\gamma} \\
&= \frac{2\gamma(L(\xi_0) - \min_\xi \ L(\xi))}{T} + C^{\frac{1}{2}}\cdot\frac{p}{n}.
\end{aligned}
$$

$\mathbb{E}\min_{t\leq T}\ \|\mathbf{b}(\xi_t)\| \leq \epsilon$ 를 만족하는 ϵ-정상점(stationary point)을 구하기 위해 다음과 같은 작업이 필요하다.

$$
\frac{1}{C^{\frac{1}{2}}}\mathbb{E}\min_{t\leq T}\ \|\mathbf{b}(\xi_t)\|^2 \leq \mathbb{E}\min_{t\leq T}\ \left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2 \leq \frac{2\gamma(L(\xi_0) - \min_\xi \ L(\xi))}{T} + C^{\frac{1}{2}}\cdot\frac{p}{n} \leq \frac{\epsilon^2}{C^{\frac{1}{2}}}.
$$

따라서 아래의 T, n을 선택하면

$$
T = \frac{C^{\frac{1}{2}}[2\gamma(L(\xi_0) - \min_\xi \ L(\xi)) + C^{\frac{1}{2}}\cdot p\delta^2]}{\epsilon^2},\ n = \frac{T}{\delta^2}
$$

ϵ-정상성(stationarity)이 만족한다.

$$
\left(\mathbb{E}\min_{t\leq T}\ \|\mathbf{b}(\xi_t)\|\right)^2 \leq \mathbb{E}\min_{t\leq T}\ \|\mathbf{b}(\xi_t)\|^2 \leq \epsilon^2.
$$

그리고 Theorem 1에 의해 두 확률 과정의 총 변동 거리에 대한 상계를 δ에 대해 표현할

수 있다.

$$D_{\mathrm{TV}}(\mu(\theta_t, t \in [T]),\ \mu(\xi_t, t \in [T])) \le O\left(\sqrt{\frac{T}{n}}\right) \le C\sqrt{\frac{T}{n}} \le O_\delta(\delta).$$

여기서 독립적인 표본의 총 개수는 $N = nT = O_{\epsilon,\delta}(\epsilon^{-4}\delta^{-2})$ 이다. □

3.2 Strongly convex case

목적함수 $L(\theta) = \mathbb{E}_{\mathbf{z}\sim P}\left[\ell(\theta, \mathbf{z})\right]$는 강볼록 함수, $\mathbf{b}(\theta) = \nabla L(\theta)$, $\mathbf{H}(\theta) = \nabla^2 L(\theta)$ 그리고 $\mathbf{V}(\theta) = \{\mathrm{Cov}\left[\nabla_\theta \ell(\theta, \mathbf{z})\right]\}^{1/2} \in \mathbb{R}^{p\times p}$가 양의 정부호 행렬이라고 하자. 그리고 아래와 같이 α, γ를 정의하자.

$$\alpha := \min_{v,w}\ \lambda_{\min}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right) > 0,$$
$$\gamma := \max_{v,w}\ \lambda_{\max}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right) > 0.$$

이때 다음의 보조정리와 정리가 성립한다.

Lemma 2. ***(Convergence: noiseless)*** *([4]의 Lemma B.1)*

함수 $L(w) : \mathbb{R}^p \to \mathbb{R}$가 매끄러운 볼록 함수라고 하자. 그리고 $\mathbf{b}(w)$, $\mathbf{H}(w)$, $\mathbf{V}(w)$, α, γ는 위에서 정의한 것이라고 할 때, 단계-크기가 $\eta = 1/\gamma$인 결정론적 갱신(deterministic update) $w_{t+1} = w_t - \eta\mathbf{V}(w_t)^{-1}\mathbf{b}(w_t)$는 다음을 만족한다.

$$L(w_{t+1}) - \min_w\ L(w) \le \left(1 - \frac{\alpha}{\gamma}\right)\left(L(w_t) - \min_w\ L(w)\right).$$

Proof. 테일러 정리에 의해, $\tilde{c} \in [0, 1]$ 에 대해 다음 식이 성립한다.

$$L(w_{t+1}) = L(w_t) + \langle \mathbf{b}(w_t), w_{t+1} - w_t\rangle + \frac{1}{2}(w_{t+1} - w_t)^T\mathbf{H}(\tilde{c}w_t + (1-\tilde{c})w_{t+1})(w_{t+1} - w_t). \quad (8)$$

편의상 $\mathbf{H}_{\tilde{c}} = \mathbf{H}(\tilde{c}w_t + (1-\tilde{c})w_{t+1})$ 라고 하자. 그리고 (8)에 $w_{t+1} = w_t - \eta\mathbf{V}(w_t)^{-1}\mathbf{b}(w_t)$를 대입한 뒤 정리하면 다음과 같다.

$$
\begin{aligned}
L(w_{t+1}) &= L(w_t) - \eta \left\langle \mathbf{b}(w_t), \mathbf{V}(w_t)^{-1}\mathbf{b}(w_t) \right\rangle + \frac{\eta^2}{2} \left[\mathbf{V}(w_t)^{-1}\mathbf{b}(w_t)\right]^T \mathbf{H}_{\tilde{c}} \left[\mathbf{V}(w_t)^{-1}\mathbf{b}(w_t)\right] \\
&= L(w_t) - \eta \left\langle \mathbf{b}(w_t), \mathbf{V}(w_t)^{-1}\mathbf{b}(w_t) \right\rangle \\
&\quad + \frac{\eta^2}{2} \left[\mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t)\right]^T \mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{H}_{\tilde{c}}\mathbf{V}(w_t)^{-\frac{1}{2}} \left[\mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t)\right] \\
&\overset{(*)}{\leq} L(w_t) - \eta \left\|\mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t)\right\|^2 + \frac{\eta^2\gamma}{2} \left\|\mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t)\right\|^2 \\
&= L(w_t) - \frac{1}{2\gamma} \left\|\mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t)\right\|^2 . \qquad (9)
\end{aligned}
$$

$(*)$에는 γ의 정의를 사용했다. 즉, $\gamma := \max_{v,w} \lambda_{\max}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right)$.

w_{t+1} 대신 임의의 w에 대해 다시 테일러 정리를 이용하면, $\tilde{c} \in [0,1]$에 대해 다음 식이 성립한다. 편의상 $\mathbf{H}_{\tilde{c}} = \mathbf{H}(\tilde{c}w_t + (1-\tilde{c})w)$라고 하자.

$$
\begin{aligned}
L(w) &= L(w_t) + \langle \mathbf{b}(w_t), w - w_t \rangle + \frac{1}{2}(w-w_t)^T\mathbf{H}_{\tilde{c}}(w-w_t) \\
&= L(w_t) + \left\langle \mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t), \mathbf{V}(w_t)^{\frac{1}{2}}(w-w_t) \right\rangle + \frac{1}{2}(w-w_t)^T\mathbf{H}_{\tilde{c}}(w-w_t) \\
&= L(w_t) + \left\langle \mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t), \mathbf{V}(w_t)^{\frac{1}{2}}(w-w_t) \right\rangle \\
&\quad + \frac{1}{2} \left[\mathbf{V}(w_t)^{\frac{1}{2}}(w-w_t)\right]^T \mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{H}_{\tilde{c}}\mathbf{V}(w_t)^{-\frac{1}{2}} \left[\mathbf{V}(w_t)^{\frac{1}{2}}(w-w_t)\right] \\
&\overset{(**)}{\geq} L(w_t) + \left\langle \mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t), \mathbf{V}(w_t)^{\frac{1}{2}}(w-w_t) \right\rangle + \frac{\alpha}{2} \left\|\mathbf{V}(w_t)^{\frac{1}{2}}(w-w_t)\right\|^2 \\
&\overset{(***)}{\geq} L(w_t) - \frac{1}{2\alpha} \left\|\mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t)\right\|^2 . \qquad (10)
\end{aligned}
$$

위 부등식의 (**)와 (***)는 아래의 식을 이용했다.

$$
\begin{aligned}
&(**) \quad \alpha := \min_{v,w} \lambda_{\min}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right) \\
&(***) \quad \left\langle \mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t), \mathbf{V}(w_t)^{\frac{1}{2}}(w-w_t) \right\rangle + \frac{\alpha}{2} \left\|\mathbf{V}(w_t)^{\frac{1}{2}}(w-w_t)\right\|^2 \geq -\frac{1}{2\alpha} \left\|\mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t)\right\|^2 \\
&\iff \left\|\frac{1}{\sqrt{\alpha}}\mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t) + \sqrt{\alpha}\mathbf{V}(w_t)^{\frac{1}{2}}(w-w_t)\right\|^2 \geq 0
\end{aligned}
$$

(9), (10)을 정리하면 다음과 같다.

$$
\begin{aligned}
&\frac{\alpha}{\gamma}(L(w) - L(w_t)) \geq -\frac{1}{2\gamma} \left\|\mathbf{V}(w_t)^{-\frac{1}{2}}\mathbf{b}(w_t)\right\|^2 \geq L(w_{t+1}) - L(w_t) \\
&\Rightarrow L(w_{t+1}) - L(w) \leq \left(1 - \frac{\alpha}{\gamma}\right)(L(w_t) - L(w)). \qquad (11)
\end{aligned}
$$

(11)에서 L의 최솟값을 가지는 w를 선택하면 다음의 결과가 나온다.

$$L(w_{t+1}) - \min_w L(w) \le \left(1 - \frac{\alpha}{\gamma}\right)\left(L(w_t) - \min_w L(w)\right).$$

□

Theorem 3. ***(MasGrad: strongly convex)*** *([4]의 Theorem 4.1)*

α, γ는 위에서 정의한 것이라고 하자. 그리고 단계-크기 $\eta = 1/\gamma$인 MasGrad 갱신 θ_t와 θ_t에 대응하는 diff_MasGrad 갱신 ξ_t를 고려하자. 그러면 임의의 정밀도(precision) $\epsilon > 0$에 대해

$$(1)\ D_{TV}(\mu(\theta_t, t \in [T]), \mu(\xi_t, t \in [T])) \le O_\epsilon\left(\sqrt{\epsilon \log 1/\epsilon}\right),$$

$$(2)\ t \le T \text{에 대해 } \mathbb{E}\left[L(\theta_t)\right] - \min_\theta\ L(\theta) \le \epsilon \text{ 와 } \mathbb{E}\left[L(\xi_t)\right] - \min_\xi\ L(\xi) \le \epsilon$$

를 만족하는 T, n은 다음과 같다:

$$T = \frac{\gamma}{\alpha}\log\frac{2(L(\xi_0) - \min_\xi\ L(\xi))}{\epsilon},\ n = \frac{4p\max_\xi\ \|\mathbf{V}(\xi)\|}{\alpha\epsilon}.$$

그리고 독립적인 표본의 총 개수는 $O_\epsilon(\epsilon^{-1}\log 1/\epsilon)$이다.

Proof. 편의상 $\mathbf{H}_{\tilde{c}} = \mathbf{H}(\tilde{c}\xi_t + (1-\tilde{c})\xi_{t+1})$ 라고 하자. 테일러 정리에 의해, $\tilde{c} \in [0,1]$에 대해 다음 식이 성립한다.

$$\begin{aligned}
&\mathbb{E}[L(\xi_{t+1})|\xi_t] \\
&= \mathbb{E}\left[L(\xi_t) + \langle \mathbf{b}(\xi_t), \xi_{t+1} - \xi_t\rangle + \frac{1}{2}(\xi_{t+1} - \xi_t)^T \mathbf{H}(\tilde{c}\xi_t + (1-\tilde{c})\xi_{t+1})(\xi_{t+1} - \xi_t)\middle|\xi_t\right] \\
&\overset{(*)}{=} \mathbb{E}\left[L(\xi_t) + \left\langle \mathbf{b}(\xi_t), -\eta\mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\ \mathbf{g}_t\right\rangle\right. \\
&\quad \left. + \frac{1}{2}\left(\eta\mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\ \mathbf{g}_t\right)^T \mathbf{H}_{\tilde{c}}\left(\eta\mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\ \mathbf{g}_t\right)\middle|\xi_t\right] \\
&= L(\xi_t) - \eta\left\langle \mathbf{b}(\xi_t), \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t)\right\rangle \\
&\quad + \mathbb{E}\left[\frac{1}{2}\left(\eta\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t) - \sqrt{2\beta^{-1}\eta}\mathbf{V}(\xi_t)^{\frac{1}{2}}\mathbf{g}_t\right)^T\right. \\
&\quad \left.\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{H}_{\tilde{c}}\mathbf{V}(\xi_t)^{-\frac{1}{2}}\left(\eta\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t) - \sqrt{2\beta^{-1}\eta}\mathbf{V}(\xi_t)^{\frac{1}{2}}\mathbf{g}_t\right)\middle|\xi_t\right] \\
&\overset{(**)}{\le} L(\xi_t) - \eta\left\langle \mathbf{b}(\xi_t), \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t)\right\rangle + \mathbb{E}\left[\frac{\gamma}{2}\left\|\eta\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t) - \sqrt{2\beta^{-1}\eta}\mathbf{V}(\xi_t)^{\frac{1}{2}}\mathbf{g}_t\right\|^2\middle|\xi_t\right] \\
&\overset{(***)}{\le} L(\xi_t) - \eta\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2 + \frac{\eta^2\gamma}{2}\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2 + \beta^{-1}\eta\gamma\mathbb{E}\left[\left\|\mathbf{V}(\xi_t)^{\frac{1}{2}}\mathbf{g}_t\right\|^2\middle|\xi_t\right] \\
&= L(\xi_t) - \frac{1}{2\gamma}\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2 + \beta^{-1}\left\langle \mathbf{I}_p, \mathbf{V}(\xi_t)\right\rangle. \qquad (12)
\end{aligned}$$

위 부등식의 (*)와 (**), (***)는 아래의 식을 이용했다.

$(*)\ \xi_{t+1} - \xi_t = -\eta \mathbf{V}(\xi_t)^{-1}\mathbf{b}(\xi_t) + \sqrt{2\beta^{-1}\eta}\ \mathbf{g}_t$

$(**)\ \gamma = \max_{v,w} \lambda_{\max}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right)$

$(***)$

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\gamma}{2}\left\|\eta\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t) - \sqrt{2\beta^{-1}\eta}\mathbf{V}(\xi_t)^{\frac{1}{2}}\mathbf{g}_t\right\|^2\middle|\xi_t\right] \\
&= \frac{\eta^2\gamma}{2}\mathbb{E}\left[\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2\middle|\xi_t\right] + \beta^{-1}\eta\gamma\mathbb{E}\left[\left\|\mathbf{V}(\xi_t)^{\frac{1}{2}}\mathbf{g}_t\right\|^2\middle|\xi_t\right] - \eta\gamma\sqrt{2\beta^{-1}\eta}\mathbb{E}\left[\mathbf{b}(\xi_t)^T\mathbf{g}_t\middle|\xi_t\right] \\
&= \frac{\eta^2\gamma}{2}\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2 + \beta^{-1}\eta\gamma\mathbb{E}\left[\left\|\mathbf{V}(\xi_t)^{\frac{1}{2}}\mathbf{g}_t\right\|^2\middle|\xi_t\right]
\end{aligned}
$$

(12)의 양변에 ξ_t에 대한 기댓값을 취한 결과와 Lemma 2의 결과로부터 다음을 얻을 수 있다:

$$
\begin{aligned}
\mathbb{E}\left[L(\xi_{t+1})\right] &\le \mathbb{E}\left[L(\xi_t)\right] - \frac{1}{2\gamma}\mathbb{E}\left[\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2\right] + \beta^{-1}\left\langle\mathbf{I}_p, \mathbb{E}\mathbf{V}(\xi_t)\right\rangle, \\
L(\xi) &\ge \mathbb{E}\left[L(\xi_t)\right] - \frac{1}{2\alpha}\mathbb{E}\left[\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \frac{\alpha}{\gamma}(L(\xi) - \mathbb{E}\left[L(\xi_t)\right]) &\ge -\frac{1}{2\gamma}\mathbb{E}\left[\left\|\mathbf{V}(\xi_t)^{-\frac{1}{2}}\mathbf{b}(\xi_t)\right\|^2\right] \\
&\ge \mathbb{E}\left[L(\xi_{t+1})\right] - \mathbb{E}\left[L(\xi_t)\right] - \beta^{-1}\left\langle\mathbf{I}_p, \mathbb{E}\mathbf{V}(\xi_t)\right\rangle \\
\Rightarrow \mathbb{E}\left[L(\xi_{t+1})\right] - L(\xi) &\le \left(1 - \frac{\alpha}{\gamma}\right)\left(\mathbb{E}\left[L(\xi_t)\right] - L(\xi)\right) + \beta^{-1}\left\langle\mathbf{I}_p, \mathbb{E}\mathbf{V}(\xi_t)\right\rangle. \qquad (13)
\end{aligned}
$$

(13)에서 L의 최솟값을 가지는 ξ를 선택하고, $\langle\mathbf{I}_p, \mathbb{E}\mathbf{V}(\xi_t)\rangle \ge p\cdot\max_\xi \|\mathbf{V}(\xi)\|$라는 사실을 이용하면 다음의 결과가 나온다:

$$
\begin{aligned}
\mathbb{E}\left[L(\xi_{t+1})\right] - \min_\xi\ L(\xi) &\le \left(1 - \frac{\alpha}{\gamma}\right)\left(\mathbb{E}\left[L(\xi_t)\right] - \min_\xi\ L(\xi)\right) + \beta^{-1}\left\langle\mathbf{I}_p, \mathbb{E}\mathbf{V}(\xi_t)\right\rangle \\
&\le \left(1 - \frac{\alpha}{\gamma}\right)\left(\mathbb{E}\left[L(\xi_t)\right] - \min_\xi\ L(\xi)\right) + \beta^{-1}p\cdot\max_\xi\ \|\mathbf{V}(\xi)\| \\
\Rightarrow \mathbb{E}\left[L(\xi_k)\right] - \min_\xi\ L(\xi) &\le \left(1 - \frac{\alpha}{\gamma}\right)^k\left(\mathbb{E}\left[L(\xi_0)\right] - \min_\xi\ L(\xi)\right) + \frac{\beta^{-1}p\cdot\max_\xi\ \|\mathbf{V}(\xi)\|}{1 - (1 - \frac{\alpha}{\gamma})}.
\end{aligned}
\qquad (14)
$$

아래의 T, n을 선택하면 (14)의 우변의 두 항 각각이 $\epsilon/2$보다 작게 되므로 $t \le T$에 대해 $\mathbb{E}\left[L(\xi_t)\right] - \min_\xi\ L(\xi) \le \epsilon$이 성립한다.

$$T = \frac{\gamma}{\alpha} \log \frac{2(L(\xi_0) - \min_\xi \ L(\xi))}{\epsilon}, \ n = \frac{4p \max_\xi \ \|\mathbf{V}(\xi)\|}{\alpha\epsilon}.$$

그리고 Theorem 1에 의해 두 확률 과정의 총 변동 거리에 대한 상계를 ϵ에 대해 표현할 수 있다.

$$D_{\mathrm{TV}}(\mu(\theta_t, t \in [T]), \ \mu(\xi_t, t \in [T])) \leq O\left(\sqrt{\frac{T}{n}}\right) \leq O_\epsilon\left(\sqrt{\epsilon \log 1/\epsilon}\right).$$

여기서 독립적인 표본의 총 개수는 $N = nT = O_\epsilon(\epsilon^{-1} \log 1/\epsilon)$ 이다. □

3.3 일반화 선형 모형에서의 가속

Definition. $L(\theta)$가 강볼록 함수라고 하자. 경사하강법의 수렴 속도는 $L(\theta)$의 헤시안 행렬(Hessian matrix)의 조건수(condition number)에 비례한다. 즉,

$$\kappa_{\mathrm{GD}} := \frac{\max_v \ \lambda_{\max}(\mathbf{H}(v))}{\min_v \ \lambda_{\min}(\mathbf{H}(v))}.$$

Theorem 3의 식 (14)에서 알 수 있듯이, MasGrad의 수렴성은 $\frac{\gamma}{\alpha}$에 의존한다:

$$\mathbb{E}\left[L(\xi_t)\right] - \min_\xi \ L(\xi) \leq \left(1 - \frac{\alpha}{\gamma}\right)^t \left(\mathbb{E}\left[L(\xi_0)\right] - \min_\xi \ L(\xi)\right) + \max_\xi \ \|\mathbf{V}(\xi)\| \cdot \frac{\gamma}{\alpha}\beta^{-1}p.$$

따라서, MasGrad의 조건수는 다음과 같이 정의할 수 있다.

$$\kappa_{\mathrm{MasGrad}} := \frac{\max_{v,w} \ \lambda_{\max}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right)}{\min_{v,w} \ \lambda_{\min}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right)} = \frac{\gamma}{\alpha}$$

일반화 선형 모형에서 MasGrad가 네스테로프 가속을 만족함을 보이자. 우선, 일반화 선형 모형을 다음과 같이 정의하자.

$$f(y; \theta, \phi) = b(y, \phi) \exp\left(\frac{y\theta - c(\theta)}{d(\phi)}\right)$$

여기서 ϕ는 알려진 모수, $\mu = \mathbb{E}\left[\mathbf{y}|\mathbf{x} = x\right] = c'(\theta)$, $c''(\theta) > 0$, $\theta = \theta(\mu) = x^T w$ 일 때, 목적함수로 음의 로그 가능도 함수를 사용하면 다음이 성립한다:

$$\ell(w, (\mathbf{x}, \mathbf{y})) = -\mathbf{y}\mathbf{x}^T w + c(\mathbf{x}^T w) \Rightarrow L(w) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P}\left[-\mathbf{y}\mathbf{x}^T w + c(\mathbf{x}^T w)\right].$$

그리고 실제 회귀 함수를 $m_*(\mathbf{x})$, 조건부 분산을 $\xi(x) = \mathbb{V}\mathrm{ar}(\mathbf{y}|\mathbf{x} = x) \in \mathbb{R}$, 편향을 $\beta(\mathbf{x}, w) = c'(\mathbf{x}^T w) - m_*(\mathbf{x}) \in \mathbb{R}$이라고 하자. 위와 같은 일반화 선형 모형에 대해서 다음의

가속이 성립한다.

Theorem 4. ***(Acceleration)*** *([4]의 Theorem 4.2)*

임의의 v, w, x에 대해, 아래 부등식을 만족하는 상수 $C > 1$가 존재한다고 가정하자.

$$0 < \max\left\{\frac{\xi(x)^2 + \beta(x,w)^2}{c''(x^T v)},\ \frac{c''(x^T v)}{\xi(x)^2}\right\} < C^{1/3}.$$

그러면 상기한 조건을 만족하는 일반화 선형 모형에 대해, 다음의 부등식이 성립한다.

$$\kappa_{MasGrad} < C\sqrt{\kappa_{GD}}.$$

Proof. 우선 $\mathrm{Cov}\left[\beta(\mathbf{x},w)\mathbf{x}\right] \preceq \mathbb{E}\left[\beta(\mathbf{x},w)^2\mathbf{x}\mathbf{x}^T\right]$이 성립한다. 그리고 편향-분산 분해를 이용하면 아래와 같이 $\mathbf{V}(w)$, $\mathbf{H}(w)$를 구할 수 있다.

$$\begin{aligned}
\mathbf{V}(w) &= \left(\mathrm{Cov}\left[\nabla\ell(w,(\mathbf{x},\mathbf{y}))\right]\right)^{1/2} = \left(\mathrm{Cov}\left[-\mathbf{y}\mathbf{x} + c'(\mathbf{x}^T w)\mathbf{x}\right]\right)^{1/2} \\
&\overset{(*)}{=} \left(\mathbb{E}\left[\xi(\mathbf{x})^2\mathbf{x}\mathbf{x}^T\right] + \mathrm{Cov}\left[\beta(\mathbf{x},w)\mathbf{x}\right]\right)^{1/2}, \qquad (15) \\
\mathbf{H}(w) &= \nabla^2 L(w) = \mathbb{E}\left[c''(\mathbf{x}^T w)\mathbf{x}\mathbf{x}^T\right]. \qquad (16)
\end{aligned}$$

위 부등식의 (*)는 아래의 편향-분산 분해를 이용했다.

$$\begin{aligned}
\xi(\mathbf{x}) = \mathbb{V}\mathrm{ar}(\mathbf{y}|\mathbf{x}) &= \mathbb{E}\left[(\mathbf{y} - \mathbb{E}\left[\mathbf{y}|\mathbf{x}\right])^2|\mathbf{x}\right] = \mathbb{E}\left[(\mathbf{y} - m_*(\mathbf{x}))^2|\mathbf{x}\right] \\
\mathrm{Cov}\left[-\mathbf{y}\mathbf{x} + c'(\mathbf{x}^T w)\mathbf{x}\right] &= \mathrm{Cov}\left[\beta(\mathbf{x},w)\mathbf{x} - (\mathbf{y} - m_*(\mathbf{x}))\mathbf{x}\right] \\
&= \mathbb{E}\left[\mathrm{Cov}\left[\beta(\mathbf{x},w)\mathbf{x} - (\mathbf{y} - m_*(\mathbf{x}))\mathbf{x}|\mathbf{x}\right]\right] \\
&\quad + \mathrm{Cov}\left[\mathbb{E}\left[\beta(\mathbf{x},w)\mathbf{x} - (\mathbf{y} - m_*(\mathbf{x}))\mathbf{x}|\mathbf{x}\right]\right] \quad \text{(by Law of total variance)} \\
&= \mathbb{E}\left[\mathrm{Cov}\left[(\mathbf{y} - m_*(\mathbf{x}))\mathbf{x}|\mathbf{x}\right]\right] + \mathrm{Cov}\left[\beta(\mathbf{x},w)\mathbf{x} - \mathbb{E}\left[(\mathbf{y} - m_*(\mathbf{x}))\mathbf{x}|\mathbf{x}\right]\right] \\
&\overset{(**)}{=} \mathbb{E}\left[\xi(\mathbf{x})\mathbf{x}\mathbf{x}^T\right] + \mathrm{Cov}\left[\beta(\mathbf{x},w)\mathbf{x}\right]
\end{aligned}$$

$$\begin{aligned}
(**)\ \mathbb{E}\left[(\mathbf{y} - m_*(\mathbf{x}))\mathbf{x}|\mathbf{x}\right] &= \mathbf{x}\left(\mathbb{E}\left[\mathbf{y}|\mathbf{x}\right] - m_*(\mathbf{x})\right) = 0 \\
\mathrm{Cov}\left[(\mathbf{y} - m_*(\mathbf{x}))\mathbf{x}|\mathbf{x}\right] &= \mathbb{E}\left[(\mathbf{y} - m_*(\mathbf{x}))^2\mathbf{x}\mathbf{x}^T|\mathbf{x}\right] = \mathbb{E}\left[(\mathbf{y} - m_*(\mathbf{x}))^2|\mathbf{x}\right]\mathbf{x}\mathbf{x}^T = \xi(\mathbf{x})\mathbf{x}\mathbf{x}^T
\end{aligned}$$

그리고 식 (15)으로부터 아래의 행렬 부등식을 얻을 수 있다.

$$
\begin{aligned}
\mathbb{E}\left[\xi(\mathbf{x})^2\mathbf{x}\mathbf{x}^T\right] &\preceq \mathbf{V}(w)^2 \\
&= \mathbb{E}\left[\xi(\mathbf{x})^2\mathbf{x}\mathbf{x}^T\right] + \mathrm{Cov}\left[\beta(\mathbf{x},w)\mathbf{x}\right] \\
&\preceq \mathbb{E}\left[\left(\xi(\mathbf{x})^2 + \beta(\mathbf{x},w)^2\right)\mathbf{x}\mathbf{x}^T\right] \qquad (17)
\end{aligned}
$$

상수 C에 대한 가정에 의해 다음 부등식이 성립한다:

$$
C^{-1/3} \leq \frac{c''(x^Tv)}{\xi(x)^2 + \beta(\mathbf{x},w)^2} \leq \frac{c''(x^Tv)}{\xi(x)^2} \leq C^{1/3}. \qquad (18)
$$

그러면 식 (16)으로부터 아래의 행렬 부등식을 얻을 수 있다.

$$
\mathbf{H}(v) = \mathbb{E}\left[c''(\mathbf{x}^Tv)\mathbf{x}\mathbf{x}^T\right] = \mathbb{E}\left[\frac{c''(x^Tv)}{\xi(\mathbf{x})^2}\xi(\mathbf{x})^2\mathbf{x}\mathbf{x}^T\right] \overset{(a),(b)}{\prec} C^{1/3}\mathbf{V}(w)^2 \qquad (19)
$$

$$
\begin{aligned}
\mathbf{H}(v) = \mathbb{E}\left[c''(\mathbf{x}^Tv)\mathbf{x}\mathbf{x}^T\right] &= \mathbb{E}\left[\frac{c''(x^Tv)}{\xi(\mathbf{x})^2 + \beta(\mathbf{x},w)^2}(\xi(\mathbf{x})^2 + \beta(\mathbf{x},w)^2)\mathbf{x}\mathbf{x}^T\right] \\
&\overset{(17),(18)}{\succ} C^{-1/3}\mathbf{V}(w)^2 \qquad (20)
\end{aligned}
$$

v가 A의 첫 번째 고유 벡터라고 하면 다음이 성립한다:

$$
A \preceq B \Rightarrow \lambda_{\max}(A) = v^TAv < v^TBv \leq \lambda_{\max}(B). \qquad (21)
$$

비슷한 방법으로

$$
A \preceq B \Rightarrow \lambda_{\min}(A) < \lambda_{\min}(B)
$$

이다. 그리고 임의의 대칭 행렬 S에 대해,

$$
A \preceq B \Rightarrow SAS \preceq SBS \qquad (22)
$$

가 성립한다.

따라서, 위 결과들을 이용해 다음 부등식을 얻을 수 있다.

$$
\begin{aligned}
&\mathbf{H}(v) \preceq C^{1/3}\mathbf{V}(w)^2 \\
&\overset{(22)}{\Rightarrow} \mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}} \preceq C^{1/3}\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{V}(w)^2\mathbf{V}(w)^{-\frac{1}{2}} \\
&\overset{(21)}{\Rightarrow} \lambda_{\max}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right) < C^{1/3}\lambda_{\max}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{V}(w)^2\mathbf{V}(w)^{-\frac{1}{2}}\right) \leq C^{1/3}\max_{w}\ \lambda_{\max}(\mathbf{V}(w)). \qquad (23)
\end{aligned}
$$

비슷한 방법으로 다음 부등식을 얻을 수 있다.

$$\lambda_{\min}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right) > C^{-1/3}\lambda_{\min}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{V}(w)^{2}\mathbf{V}(w)^{-\frac{1}{2}}\right) \geq C^{-1/3}\min_{w}\ \lambda_{\min}(\mathbf{V}(w)). \tag{24}$$

κ_{MasGrad}의 정의와 위 결과들을 이용해 최종적인 부등식을 얻을 수 있다.

$$\begin{aligned}
\kappa_{\text{MasGrad}} &= \frac{\max_{v,w}\ \lambda_{\max}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right)}{\min_{v,w}\ \lambda_{\min}\left(\mathbf{V}(w)^{-\frac{1}{2}}\mathbf{H}(v)\mathbf{V}(w)^{-\frac{1}{2}}\right)} \\
&\overset{(23),(24)}{\leq} \frac{C^{1/3}\max_{w}\ \lambda_{\max}(\mathbf{V}(w))}{C^{-1/3}\min_{w}\ \lambda_{\min}(\mathbf{V}(w))} \\
&\leq C^{2/3}\sqrt{\frac{\max_{w}\ \lambda_{\max}(\mathbf{V}(w)^2)}{\min_{w}\ \lambda_{\min}(\mathbf{V}(w)^2)}} \\
&\overset{(19),(20)}{\leq} C\sqrt{\frac{\max_{v}\ \lambda_{\max}(\mathbf{H}(v))}{\min_{v}\ \lambda_{\min}(\mathbf{H}(v))}} \\
&= C\sqrt{\kappa_{\text{GD}}}
\end{aligned}$$

□

Remark. Theorem 4와 Lemma 2의 결과로부터 노이즈가 없는 환경(noiseless setting)에서 GD의 시간 복잡도는 $O(\kappa_{\text{GD}}\log(1/\epsilon))$인 반면, MasGrad의 시간 복잡도는 $O(\sqrt{\kappa_{\text{GD}}}\log(1/\epsilon))$임을 알 수 있다. 특히 GD의 조건수($\kappa_{\text{GD}}$)가 클 때, 위 시간 복잡도의 차이는 극명하게 나타난다.

4 매끄럽지 않은 정규화 문제

이 장에서는 매끄럽지 않은 정규화 항이 있는 목적함수를 최소화하는 문제(non-smooth regularization problem)에 대해서 알아보자. 이 문제를 최적화 식으로 나타내면 다음과 같다:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; L(x) := g(x) + h(x), \tag{25}$$

여기서 g는 볼록이고 연속적으로 미분가능한 함수이고 h는 볼록이지만 연속일 필요는 없는 정규화 항이다. 대표적으로 아래의 예시가 있다.

1. Sparse regression

$$L(w) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P}\left[\frac{1}{2}\left(\mathbf{x}^T w - \mathbf{y}\right)^2\right] + \lambda \left\|w\right\|_1 := g(w) + h(w).$$

2. Low rank matrix trace regression

$$L(W) = \mathbb{E}_{(\mathbf{X},\mathbf{y})\sim P}\left[\frac{1}{2}\left(\langle \mathbf{X}, W\rangle - \mathbf{y}\right)^2\right] + \lambda \left\|W\right\|_* := g(W) + h(W).$$

정규화 항이 있는 목적함수 (25)을 최적화하는 방법으로는 정규화 항이 없는 목적함수 (1)을 최적화하는 방법인 뉴턴 방법과 MasGrad에 각각 대응하는 근위 뉴턴 방법(proximal-newton type method, ProxNewton [3])과 적률 조정 근위 경사하강법(moment adjusted proximal gradient descent method, MadProx [4])이 있다. 두 방법을 설명하기에 앞서, scaled proximal mapping을 정의하자.

Definition. h는 볼록 함수이고 H는 양의 정부호 행렬이라고 하자. 그러면 x에서 h의 scaled proximal mapping은 다음과 같다:

$$\text{prox}_h^{\mathbf{H}}(x) := \underset{y\in\mathbb{R}^n}{\text{argmin}} \;\left[\frac{1}{2}\left\|y - x\right\|_{\mathbf{H}}^2 + h(y)\right]. \tag{26}$$

Note. Scaled proximal mapping은 (unscaled) proximal mapping과 많은 성질을 공유하고 있다:

1. H가 양의 정부호 행렬이라면 근위 함수는 강볼록이므로, $\text{prox}_h^H(x)$는 $x \in \text{dom } h$에 유일하게 존재한다.

2. $\partial h(x)$가 x에서 h의 하방미분(subdifferential)이라고 하자. 그러면 $\text{prox}_h^H(x)$는 다음을 만족한다.

$$H\left(x - \text{prox}_h^H(x)\right) \in \partial h\left(\text{prox}_h^H(x)\right).$$

3. Scaled proximal mapping은 H-노름에 있어서 단단히 비확장적(firmly nonexpansive) 이다.

$$\left\|\mathrm{prox}_h^H(x) - \mathrm{prox}_h^H(y)\right\|_H \leq \|x - y\|_H .$$

4.1 ProxNewton

목적함수 (25)을 최소화하기 위한 대표적인 이차 반복법으로 근위 뉴턴 방법(Proximal-Newton type method, ProxNewton)이 있다. 참고문헌 [3]의 ProxNewton은 뉴턴 방법의 proximal 버전으로, g의 곡률을 계산하기 위해 g의 헤시안 행렬 $\nabla^2 g(x)$을 사용한다. 다시 말해서, ProxNewton의 검색 방향 Δx(식 (2)에서 $-\mathbf{h}$에 해당)는 다음과 같은 부분 문제를 풀어서 구할 수 있다:

$$\begin{aligned} \Delta x &= \underset{d}{\mathrm{argmin}}\ \hat{L}(x+d) := \hat{g}(x+d) + h(x+d), \\ \text{where } \hat{g}(y) &= g(x) + \nabla g(x)^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 g(x)(y-x). \end{aligned} \tag{27}$$

4.2 MadProx

함수 g의 이차 근사식 (27)에서 g의 헤시안 행렬 $\nabla^2 g(x)$ 대신 이차 적률 행렬 $\mathbf{V}(x) = \{\mathrm{Cov}\,[\nabla g(x)]\}^{1/2}$을 사용하여 목적함수를 최적화하는 방법을 적률 조정 근위 경사하강법(moment adjusted proximal gradient descent method, MadProx)이라고 한다. ProxNewton과 마찬가지로 MadProx의 검색 방향 Δx는 다음과 같은 부분 문제를 풀어서 구할 수 있다:

$$\begin{aligned} \Delta x &= \underset{d}{\mathrm{argmin}}\ \hat{L}(x+d) := \hat{g}(x+d) + h(x+d), \\ \text{where } \hat{g}(y) &= g(x) + \nabla g(x)^T(y-x) + \frac{1}{2}(y-x)^T \mathbf{V}(x)(y-x). \end{aligned} \tag{28}$$

5 MadProx의 성질

식 (28)의 검색 방향 Δx는 scaled proximal mapping을 사용하여 표현할 수 있기 때문에 아래의 Δx을 사용하여 적률 조정 근위 경사하강법(moment adjusted proximal gradient descent method, MadProx)을 정의할 것이다:

$$\Delta x = \mathrm{prox}_h^{\mathbf{V}(x)}\left(x - \mathbf{V}(x)^{-1}\nabla g(x)\right) - x.$$

또한, 적당히 작은 $\eta \in (0, 1]$에 대하여 검색 방향 Δx을 사용하여 x를 갱신하면, 즉 $x \leftarrow x + \eta\Delta x$, 목적함수 L은 감소한다:

$$L(x + \eta\Delta x) \leq L(x).$$

위 부등식을 만족하는 적당히 작은 단계-크기 η는 직선 탐색 과정(line search procedure)을 통해 구할 것이다.

참고문헌 [4]의 MadProx는 이차 적률 행렬 $\mathbf{V}(x_t)$이 갱신되는 변수 x_t에 의존하지 않고, 즉 $\mathbf{V}(x_t) = \mathbf{V}$ for all t, 매 반복마다 단계-크기가 $\eta = 1/\gamma$로 고정되었을 때의 MadProx 갱신 x_t의 수렴 분석을 하였다. 본 논문에서는 [3]을 참고하여 매 반복마다 고정된 η가 아닌 Algorithm 1의 직선 탐색 과정으로 구한 η_t를 이용하고 (29)에서 정의한 검색 방향 Δx를 이용하여 MadProx 갱신 x_t을 정의하였다. MadProx 갱신 x_t에 대한 자세한 과정은 Algorithm 2에 나와 있다.

5.1 검색 방향

Proposition 2. *MadProx의 검색 방향 Δx는 scaled proximal mapping을 사용해서 정리하면 다음과 같다:*

$$\Delta x = \mathrm{prox}_h^{\mathbf{V}(x)}\left(x - \mathbf{V}(x)^{-1}\nabla g(x)\right) - x. \tag{29}$$

Proof. $x_+ = x + \Delta x$라고 하자. 그러면 (28)로부터

$$x_+ = \underset{y}{\mathrm{argmin}}\ \hat{L}(y) = \underset{y}{\mathrm{argmin}}\ \hat{g}(y) + h(y)$$

이다. $\hat{g}(y)$의 정의를 이용하면 다음과 같이 x_+를 scaled proximal mapping을 사용하여 표현할 수 있다:

$$
\begin{aligned}
x_+ &= \underset{y}{\operatorname{argmin}}\ \hat{L}(y) \\
&= \underset{y}{\operatorname{argmin}}\ \hat{g}(y) + h(y) \\
&= \underset{y}{\operatorname{argmin}}\ \left[g(x) + \nabla g(x)^T(y-x) + \frac{1}{2}(y-x)^T\mathbf{V}(x)(y-x) + h(y)\right] \\
&= \underset{y}{\operatorname{argmin}}\ \left[\frac{1}{2}\nabla g(x)^T\mathbf{V}(x)^{-1}\nabla g(x) + \nabla g(x)^T(y-x) + \frac{1}{2}(y-x)^T\mathbf{V}(x)(y-x) + h(y)\right] \\
&= \underset{y}{\operatorname{argmin}}\ \left[\frac{1}{2}\left\|y - x + \mathbf{V}(x)^{-1}\nabla g(x)\right\|^2_{\mathbf{V}(x)} + h(y)\right] \\
&= \operatorname{prox}_h^{\mathbf{V}(x)}\left(x - \mathbf{V}(x)^{-1}\nabla g(x)\right).
\end{aligned}
$$

$x_+ = x + \Delta x$이므로

$$
\Delta x = \operatorname{prox}_h^{\mathbf{V}(x)}\left(x - \mathbf{V}(x)^{-1}\nabla g(x)\right) - x.
$$

이다. □

Remark. 즉, MadProx 갱신 x_t는 다음과 같다:

$$
\begin{aligned}
x_{t+1} &= x_t + \eta_t \Delta x_t, \qquad (30) \\
\text{where } \Delta x_t &= \operatorname{prox}_h^{\mathbf{V}(x_t)}\left(x_t - \mathbf{V}(x_t)^{-1}\nabla g(x_t)\right) - x_t.
\end{aligned}
$$

여기서 η_t는 단계-크기(step-size)이다.

Proposition 3. *$\mathbf{V}(x)$가 양의 정부호 행렬일 때, (29)에서 정의한 MadProx의 검색 방향 Δx는 다음을 성질을 만족한다.*

$$
L(x_+) \le L(x) + \eta\left(\nabla g(x)^T\Delta x + h(x+\Delta x) - h(x)\right) + O(\eta^2), \qquad (31)
$$

$$
\nabla g(x)^T\Delta x + h(x+\Delta x) - h(x) \le -\Delta x^T\mathbf{V}(x)\Delta x. \qquad (32)
$$

여기서 $x_+ = x + \eta\Delta x$이고 η는 단계-크기이다.

Proof. 증명은 [3]의 Proposition 2.4을 참조하였다.

$\eta \in (0, 1]$에 대해

$$
\begin{aligned}
L(x_+) - L(x) &= g(x_+) - g(x) + h(x_+) - h(x) \\
&\le g(x_+) - g(x) + \eta h(x+\Delta x) + (1-\eta)h(x) - h(x) \\
&\le g(x_+) - g(x) + \eta\left(h(x+\Delta x) - h(x)\right) \\
&= \nabla g(x)^T(\eta\Delta x) + \eta\left(h(x+\Delta x) - h(x)\right) + O(\eta^2) \\
&= \eta\left(\nabla g(x)^T\Delta x + h(x+\Delta x) - h(x)\right) + O(\eta^2).
\end{aligned}
$$

위 부등식의 마지막 줄로부터 (31)이 성립함을 알 수 있다.

한편, (28)으로부터

$$
\Delta x = \underset{d}{\operatorname{argmin}}\ \hat{g}(x+d) + h(x+d) = \underset{d}{\operatorname{argmin}}\ \left[g(x) + \nabla g(x)^T d + \frac{1}{2}d^T\mathbf{V}(x)d + h(x+d)\right]
$$

이므로

$$
\begin{aligned}
&\nabla g(x)^T\Delta x + \frac{1}{2}\Delta x^T\mathbf{V}(x)\Delta x + h(x+\Delta x) \\
&\quad\le \nabla g(x)^T(\eta\Delta x) + \frac{1}{2}(\eta\Delta x)^T\mathbf{V}(x)(\eta\Delta x) + h(x+\eta\Delta x) \\
&\quad= \eta\nabla g(x)^T\Delta x + \frac{1}{2}\eta^2\Delta x^T\mathbf{V}(x)\Delta x + h(x_+) \\
&\quad\le \eta\nabla g(x)^T\Delta x + \frac{1}{2}\eta^2\Delta x^T\mathbf{V}(x)\Delta x + \eta h(x+\Delta x) + (1-\eta)h(x).
\end{aligned}
$$

위 부등식의 우변을 좌변으로 이항하면 다음과 같이 정리할 수 있다:

$$
\begin{aligned}
&(1-\eta)\nabla g(x)^T\Delta x + \frac{1}{2}(1-\eta^2)\Delta x^T\mathbf{V}(x)\Delta x + (1-\eta)(h(x+\Delta x) - h(x)) \le 0 \\
&\overset{\div\,(1-\eta)}{\Rightarrow}\ \nabla g(x)^T\Delta x + \frac{1}{2}(1+\eta)\Delta x^T\mathbf{V}(x)\Delta x + h(x+\Delta x) - h(x) \le 0 \\
&\Rightarrow\ \nabla g(x)^T\Delta x + h(x+\Delta x) - h(x) \le -\frac{1}{2}(1+\eta)\Delta x^T\mathbf{V}(x)\Delta x. \qquad (33)
\end{aligned}
$$

(33)에서 $\eta \to 1$일 때, 부등식 (32)이 성립한다. □

Note. (31)에 (32)을 대입하면 $L(x_+) \le L(x) - \eta\Delta x^T\mathbf{V}(x)\Delta x + O(\eta^2)$이고 $\mathbf{V}(x)$는 양의 정부호 행렬, 즉 $\Delta x^T\mathbf{V}(x)\Delta x > 0$, 이므로 적당히 작은 η에서 검색 방향 Δx을 사용하여 x를 갱신하면 목적함수 L은 감소한다:

$$
L(x_+) \le L(x) - \eta\Delta x^T\mathbf{V}(x)\Delta x + O(\eta^2) \le L(x).
$$

Proposition 4. *$\mathbf{V}(x)$가 양의 정부호 행렬일 때, $x^\star$가 $L(x)$의 최적해인 것과 $x^\star$에서의 검색 방향 Δx이 0인 것은 동치이다.*

Proof. 증명은 [3]의 Proposition 2.5를 참조하였다.

$x^\star$에서의 검색 방향 Δx을 $\Delta x^\star$라고 표기하자.

($\Rightarrow$) $\Delta x^\star \neq 0$이라면, $\Delta x^\star$는 $x^\star$에서 L이 감소하는 방향이다. 즉, $x^\star$는 $L(x)$의 최적해가 아니다. 따라서, $x^\star$가 $L(x)$의 최적해라면 $\Delta x^\star = 0$이다.

($\Leftarrow$) $\Delta x^\star = 0$이라고 하자. 그러면 $\Delta x^\star = \mathrm{argmin}_d \ \hat{L}^\star(x^\star + d)$이므로 모든 $\eta > 0$와 d에 대해 아래의 부등식을 만족한다.

$$\begin{aligned}
&\hat{L}^\star(x^\star + \eta d) \geq \hat{L}^\star(x^\star) \\
&\Rightarrow \ \nabla g(x^\star)^T(\eta d) + \frac{1}{2}\eta^2 d^T \mathbf{V}(x^\star) d + h(x^\star + \eta d) \geq h(x^\star) \\
&\Rightarrow \ h(x^\star + \eta d) - h(x^\star) \geq -\eta \nabla g(x^\star)^T d - \frac{1}{2}\eta^2 d^T \mathbf{V}(x^\star) d. \qquad (34)
\end{aligned}$$

$\nabla_d L(x^\star)$를 $x^\star$에서 L의 d 방향의 방향 도함수라고 하면 모든 d에 대하여 $\nabla_d L(x^\star) \geq 0$이다:

$$\begin{aligned}
\nabla_d L(x^\star) &= \lim_{\eta \to 0} \frac{L(x^\star + \eta d) - L(x^\star)}{\eta} \\
&= \lim_{\eta \to 0} \frac{g(x^\star + \eta d) - g(x^\star) + h(x^\star + \eta d) - h(x^\star)}{\eta} \\
&= \lim_{\eta \to 0} \frac{\eta \nabla g(x^\star)^T d + O(\eta^2) + h(x^\star + \eta d) - h(x^\star)}{\eta} \\
&\overset{(34)}{\geq} \lim_{\eta \to 0} \frac{\eta \nabla g(x^\star)^T d + O(\eta^2) - \eta \nabla g(x^\star)^T d - \frac{1}{2}\eta^2 d^T \mathbf{V}(x^\star) d}{\eta} \\
&= 0.
\end{aligned}$$

따라서, 모든 $\eta > 0$와 d에 대해 $L(x^\star) \leq L(x^\star + \eta d)$를 만족하고 $L(x)$는 볼록 함수이므로 $x^\star$는 $L(x)$의 최적해이다. □

5.2 직선 탐색 과정

본 논문에서는 MadProx 갱신 (30)에서 정의한 MadProx 갱신의 단계-크기 η_t를 구하기 위해 각 갱신마다 직선 탐색 과정(line search procedure)을 이용할 것이다. 직선 탐색 과정은 갱신되는 값 x_+이 다음의 충분 하강 조건(sufficient descent condition)을 만족하도록 $\eta \in (0, 1]$를 조정하는 방법이다:

$$L(x_+) \leq L(x) + \alpha\eta\lambda, \tag{35}$$
$$\text{where } \lambda := \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) \ (< 0), \text{ and } \alpha \in (0, 0.5).$$

구체적인 직선 탐색 과정은 Algorithm 1에 나와 있다. 한편, 적절한 조건하에서 직선 탐색 과정을 통해 η를 구했다면 η의 범위는 아래의 Lemma 3으로부터 $\eta \leq \min\left\{1,\ \frac{2m}{L_1}(1-\alpha)\right\}$ 임을 알 수 있다.

Algorithm 1: 직선 탐색 과정

Given x_t, Δx_t.
Choose $\alpha \in (0,\ 0.5)$, $\beta \in (0,\ 1)$, $\eta = 1$.
repeat
  $\eta \leftarrow \beta\eta$.
  $x_{t+1} \leftarrow x_t + \eta\Delta x_t$.
until $L(x_{t+1}) \leq L(x_t) + \alpha\eta \cdot \big(\nabla g(x_t)^T \Delta x_t + h(x + \Delta x_t) - h(x_t)\big)$.
Terminate with $\eta_t = \eta$.

Lemma 3. *모든 x에 대해 $\mathbf{V}(x) \succeq mI$를 만족하는 상수 $m > 0$이 존재하고, $\nabla g(x)$는 L_1-립시츠 연속이라고 하자. 그러면 단계-크기 η가 아래와 같을 때 충분 하강 조건 (35)을 만족한다.*

$$\eta \leq \min\left\{1,\ \frac{2m}{L_1}(1-\alpha)\right\}$$

.

Proof. 증명은 [3]의 Lemma 2.6를 참조하였다.

$$
\begin{aligned}
&L(x_+) - L(x) \\
&\quad = g(x_+) - g(x) + h(x_+) - h(x) \\
&\quad \le \int_0^1 \nabla g(x + s(\eta\Delta x))^T (\eta\Delta x)\ ds + \eta h(x+\Delta x) + (1-\eta)h(x) - h(x) \\
&\quad = \nabla g(x)^T(\eta\Delta x) + \eta(h(x+\Delta x) - h(x)) + \int_0^1 (\nabla g(x+s(\eta\Delta x)) - \nabla g(x))^T(\eta\Delta x)\ ds \\
&\quad \le \eta\left(\nabla g(x)^T\Delta x + h(x+\Delta x) - h(x) + \int_0^1 \|\nabla g(x+s(\eta\Delta x)) - \nabla g(x)\|\,\|\Delta x\|\ ds\right) \\
&\quad \le \eta\left(\nabla g(x)^T\Delta x + h(x+\Delta x) - h(x) + \int_0^1 L_1\,\|x+s(\eta\Delta x) - x\|\,\|\Delta x\|\ ds\right) \\
&\quad \le \eta\left(\nabla g(x)^T\Delta x + h(x+\Delta x) - h(x) + \frac{L_1\eta}{2}\|\Delta x\|^2\right) \\
&\quad = \eta\left(\lambda + \frac{L_1\eta}{2}\|\Delta x\|^2\right). \qquad (36)
\end{aligned}
$$

단계-크기 η가 $\eta \le \frac{2m}{L_1}(1-\alpha)$을 만족할 때

$$
\frac{L_1\eta}{2}\|\Delta x\|^2 \le m(1-\alpha)\|\Delta x\|^2 \le (1-\alpha)\Delta x^T\mathbf{V}(x)\Delta x \overset{(32)}{\le} -(1-\alpha)\lambda. \qquad (37)
$$

따라서, $\eta \le \min\left\{1,\ \frac{2m}{L_1}(1-\alpha)\right\}$일 때 아래의 충분 하강 조건을 만족한다:

$$
\begin{aligned}
L(x_+) - L(x) &\overset{(36)}{\le} \eta\left(\lambda + \frac{L_1\eta}{2}\|\Delta x\|^2\right) \\
&\overset{(37)}{\le} \eta(\lambda - (1-\alpha)\lambda) \\
&= \alpha\eta\lambda.
\end{aligned}
$$

□

Algorithm 2: MadProx

Require: starting point $x_0 \in \mathrm{dom}\ L$.
repeat
Choose $\mathbf{V}(x_t)$.
Solve the subproblem for a search direction:
$\Delta x_t \leftarrow \mathrm{prox}_h^{\mathbf{V}(x_t)}\left(x_t - \mathbf{V}(x_t)^{-1}\nabla g(x_t)\right) - x_t$.
Select η_t with a bactracking line search.
Update: $x_{t+1} \leftarrow x_t + \eta_t\Delta x_t$.
until stopping conditions are satisfied.

Algorithm 3: ProxNewton

Require: starting point $x_0 \in \operatorname{dom} L$.

repeat

Choose $\nabla^2 g(x_t)$.

Solve the subproblem for a search direction:

$\Delta x_t \leftarrow \operatorname{prox}_h^{\nabla^2 g(x_t)} \left(x_t - \nabla^2 g(x_t)^{-1} \nabla g(x_t)\right) - x_t$.

Select η_t with a bactracking line search.

Update: $x_{t+1} \leftarrow x_t + \eta_t \Delta x_t$.

until stopping conditions are satisfied.

6 MadProx의 수렴성

이 장에서는 이차 적률 행렬 $\mathbf{V}(x_t)$이 갱신되는 변수 x_t에 의존하는 경우에도 x_t가 특정 값에 수렴하고, 그 수렴 속도가 q-quadratically 지역 수렴성을 가짐을 보인다.

6.1 $\mathbf{V}(x_t) = \mathbf{V}$일 때 MadProx의 수렴성

참고문헌 [4]에서는 적률 조정 근위 함수(moment adjusted proximal function)와 MadProx 갱신을 다음과 같이 정의한다.

$$\mathrm{prox}_{\eta,\mathbf{V}}(x) := \underset{y}{\mathrm{argmin}} \left[\frac{1}{2\eta}\|y-x\|_{\mathbf{V}}^2 + h(y)\right], \tag{38}$$

$$\mathrm{MadProx} : x_{t+1} = \mathrm{prox}_{\eta,\mathbf{V}}\left(x_t - \eta\mathbf{V}^{-1}\nabla g(x_t)\right). \tag{39}$$

그리고 $\mathbf{H}(x) = \nabla^2 g(x)$라고 할 때 α, γ를 다음과 같이 정의하면 :

$$\alpha := \min_x \ \lambda_{\min}\left(\mathbf{V}^{-\frac{1}{2}}\mathbf{H}(x)\mathbf{V}^{-\frac{1}{2}}\right) > 0, \ \gamma := \max_x \ \lambda_{\max}\left(\mathbf{V}^{-\frac{1}{2}}\mathbf{H}(x)\mathbf{V}^{-\frac{1}{2}}\right) > 0.$$

여기서 $\mathbf{V}(x)$는 x에 의존하지 않는 경우만을 고려한다.

Theorem 5. *([4]의 Proposition 4.1)*

식 (39)에서 정의한 MadProx 갱신 x_t의 단계-크기가 $\eta = 1/\gamma$라고 하자. 그러면 $L(x_T) - \min_x \ L(x) \le \epsilon$를 만족하는 T의 범위는 다음과 같다:

$$T \ge \frac{\gamma}{\alpha}\log\left(\frac{\alpha}{2\epsilon}\|x_0 - x^\star\|_{\mathbf{V}}^2 + 1\right),$$

여기서 $x^\star = \underset{x}{\mathrm{argmin}} \ L(x)$이다.

Proof. 임의의 x와 $z \in \partial h(x_{t+1})$를 만족하는 임의의 sub-gradient z에 대해 다음 식이 성립한다.

$$
\begin{aligned}
L(x_{t+1}) &= g(x_{t+1}) + h(x_{t+1}) \\
&\overset{(*)}{\leq} \left[g(x_t) + \langle \nabla g(x_t), x_{t+1} - x_t \rangle + \frac{1}{2} \|x_{t+1} - x_t\|^2_{\mathbf{H}(\tilde{c})} \right] + [h(x) + \langle z, x_{t+1} - x \rangle] \\
&\overset{(**)}{\leq} \left[g(x) + \langle \nabla g(x_t), x_t - x \rangle - \frac{1}{2} \|x_t - x\|^2_{\mathbf{H}(c')} \right] \\
&\quad + \left[\langle \nabla g(x_t), x_{t+1} - x_t \rangle + \frac{1}{2} \|x_{t+1} - x_t\|^2_{\mathbf{H}(\tilde{c})} \right] + [h(x) + \langle z, x_{t+1} - x \rangle] \\
&= \left[g(x) + \langle \nabla g(x_t), x_{t+1} - x \rangle + \frac{1}{2} \|x_{t+1} - x_t\|^2_{\mathbf{H}(\tilde{c})} - \frac{1}{2} \|x_t - x\|^2_{\mathbf{H}(c')} \right] \\
&\quad + [h(x) + \langle z, x_{t+1} - x \rangle] \\
&= L(x) + \langle \nabla g(x_t) + z, x_{t+1} - x \rangle + \frac{1}{2} \|x_{t+1} - x_t\|^2_{\mathbf{H}(\tilde{c})} - \frac{1}{2} \|x_t - x\|^2_{\mathbf{H}(c')} . \qquad (40)
\end{aligned}
$$

$$
\begin{aligned}
(*) \quad & g(x_{t+1}) = g(x_t) + \langle \nabla g(x_t), x_{t+1} - x_t \rangle + \frac{1}{2}(x_{t+1} - x_t)^T \mathbf{H}(\tilde{c})(x_{t+1} - x_t) && \text{(테일러 정리)}, \\
& z \in \partial h(x_{t+1}) \Rightarrow h(x) \geq h(x_{t+1}) + \langle z, x - x_{t+1} \rangle \ \text{ for all } x && \text{(sub-gradient z의 성질)}, \\
(**) \quad & g(x) = g(x_t) + \langle \nabla g(x_t), x - x_t \rangle + \frac{1}{2}(x - x_t)^T \mathbf{H}(c')(x - x_t) && \text{(테일러 정리)}.
\end{aligned}
$$

(38), (39)에서 정의한 적률 조정 근위 함수와 MadProx 갱신으로부터 다음 식을 얻을 수 있다.

$$
\begin{aligned}
& x_{t+1} = \mathrm{prox}_{\eta, \mathbf{V}} \left(x_t - \eta \mathbf{V}^{-1} \nabla g(x_t) \right) \\
& \Rightarrow x_{t+1} = \underset{u}{\mathrm{argmin}} \ \left[\frac{1}{2\eta} \left\| u - \left(x_t - \eta \mathbf{V}^{-1} \nabla g(x_t) \right) \right\|^2_{\mathbf{V}} + h(u) \right] \\
& \Rightarrow 0 \in \frac{1}{\eta} \mathbf{V} \left(x_{t+1} - x_t + \eta \mathbf{V}^{-1} \nabla g(x_t) \right) + \partial h(x_{t+1}). \qquad (41)
\end{aligned}
$$

그리고 (41)와 $z \in \partial h(x_{t+1})$로부터 다음 식을 얻는다.

$$
\nabla g(x_t) + z = \frac{1}{\eta} \mathbf{V}(x_t - x_{t+1}). \qquad (42)
$$

(42)와 α, γ의 정의를 이용해 (40)을 다음과 같이 정리할 수 있다.

$$
\begin{aligned}
L(x_{t+1}) &\le L(x) + \langle \nabla g(x_t) + z, x_{t+1} - x \rangle + \frac{1}{2}\|x_{t+1} - x_t\|^2_{\mathbf{H}(\tilde{c})} - \frac{1}{2}\|x_t - x\|^2_{\mathbf{H}(c')} \\
&\overset{(42)}{=} L(x) + \left\langle \frac{1}{\eta}\mathbf{V}(x_t - x_{t+1}), x_{t+1} - x \right\rangle + \frac{1}{2}\|x_{t+1} - x_t\|^2_{\mathbf{H}(\tilde{c})} - \frac{1}{2}\|x_t - x\|^2_{\mathbf{H}(c')} \\
&= L(x) + \left\langle \frac{1}{\eta}\mathbf{V}(x_t - x_{t+1}), x_{t+1} - x \right\rangle + \frac{1}{2}(x_{t+1} - x_t)^T\mathbf{H}(\tilde{c})(x_{t+1} - x_t) - \frac{1}{2}(x_t - x)^T\mathbf{H}(c')(x_t - x) \\
&= L(x) + \left\langle \frac{1}{\eta}\mathbf{V}(x_t - x_{t+1}), x_{t+1} - x \right\rangle \\
&\quad + \frac{1}{2}(x_{t+1} - x_t)^T\mathbf{V}^{\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}\mathbf{H}(\tilde{c})\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{\frac{1}{2}}(x_{t+1} - x_t) - \frac{1}{2}(x_t - x)^T\mathbf{V}^{\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}\mathbf{H}(c')\mathbf{V}^{-\frac{1}{2}}\mathbf{V}^{\frac{1}{2}}(x_t - x) \\
&\le L(x) + \left\langle \frac{1}{\eta}\mathbf{V}(x_t - x_{t+1}), x_{t+1} - x \right\rangle + \frac{\gamma}{2}\|x_{t+1} - x_t\|^2_{\mathbf{V}} - \frac{\alpha}{2}\|x_t - x\|^2_{\mathbf{V}}. \qquad (43)
\end{aligned}
$$

$\eta = \frac{1}{\gamma}$일 때, (43)에 $x = x_t$를 대입하면 다음과 같다.

$$
\begin{aligned}
L(x_{t+1}) &\le L(x_t) - (\frac{1}{\eta} - \frac{\gamma}{2})\|x_{t+1} - x_t\|^2_{\mathbf{V}} \\
&= L(x_t) - \frac{\gamma}{2}\|x_{t+1} - x_t\|^2_{\mathbf{V}} \le L(x_t). \qquad (44)
\end{aligned}
$$

그리고 (43)에 $x = x^\star (= \underset{x}{\operatorname{argmin}}\ L(x))$를 대입하면 다음과 같다.

$$
\begin{aligned}
L(x_{t+1}) - L(x^\star) &\le \gamma \langle x_t - x_{t+1}, x_{t+1} - x^\star \rangle_{\mathbf{V}} + \frac{\gamma}{2}\|x_{t+1} - x_t\|^2_{\mathbf{V}} \\
&\quad - \frac{\alpha}{2}\|x_t - x^\star\|^2_{\mathbf{V}} \\
&\overset{(***)}{=} -\frac{\gamma}{2}\|x_{t+1} - x^\star\|^2_{\mathbf{V}} + \frac{\gamma}{2}\|x_{t+1} - x^\star + x_t - x_{t+1}\|^2_{\mathbf{V}} \\
&\quad - \frac{\alpha}{2}\|x_t - x^\star\|^2_{\mathbf{V}} \\
&= \frac{\gamma - \alpha}{2}\|x_t - x^\star\|^2_{\mathbf{V}} - \frac{\gamma}{2}\|x_{t+1} - x^\star\|^2_{\mathbf{V}}
\end{aligned}
$$

$$
\Rightarrow \frac{2}{\gamma - \alpha}[L(x_{t+1}) - L(x^\star)] \le \|x_t - x^\star\|^2_{\mathbf{V}} - \frac{\gamma}{\gamma - \alpha}\|x_{t+1} - x^\star\|^2_{\mathbf{V}}. \qquad (45)
$$

$$
\begin{aligned}
(***) \quad & \|x_{t+1} - x^\star + x_t - x_{t+1}\|^2_{\mathbf{V}} \\
&= \|x_{t+1} - x^\star\|^2_{\mathbf{V}} + 2\langle x_t - x_{t+1}, x_{t+1} - x^\star \rangle_{\mathbf{V}} + \|x_{t+1} - x_t\|^2_{\mathbf{V}}.
\end{aligned}
$$

(44), (45)을 이용하여 아래와 같이 $\frac{2}{\alpha}\left[\left(\frac{\gamma}{\gamma-\alpha}\right)^T - 1\right](L(x_T) - L(x^\star))$에 대한 부등식을 얻을 수 있다.

$$
\begin{aligned}
&\frac{2}{\alpha}\left[\left(\frac{\gamma}{\gamma-\alpha}\right)^T - 1\right](L(x_T) - L(x^\star)) \\
&= \left[\frac{2}{\gamma-\alpha}\sum_{t=0}^{T-1}\left(\frac{\gamma}{\gamma-\alpha}\right)^t\right](L(x_T) - L(x^\star)) \\
&\overset{(44)}{\leq} \frac{2}{\gamma-\alpha}\sum_{t=0}^{T-1}\left(\frac{\gamma}{\gamma-\alpha}\right)^t (L(x_{t+1}) - L(x^\star)) \\
&\overset{(45)}{\leq} \sum_{t=0}^{T-1}\left(\frac{\gamma}{\gamma-\alpha}\right)^t \left\{\|x_t - x^\star\|_{\mathbf{V}}^2 - \frac{\gamma}{\gamma-\alpha}\|x_{t+1} - x^\star\|_{\mathbf{V}}^2\right\} \\
&\leq \sum_{t=0}^{T-1}\left\{\left(\frac{\gamma}{\gamma-\alpha}\right)^t \|x_t - x^\star\|_{\mathbf{V}}^2 - \left(\frac{\gamma}{\gamma-\alpha}\right)^{t+1}\|x_{t+1} - x^\star\|_{\mathbf{V}}^2\right\} \\
&= \|x_0 - x^\star\|_{\mathbf{V}}^2 - \left(\frac{\gamma}{\gamma-\alpha}\right)^T \|x_T - x^\star\|_{\mathbf{V}}^2 \\
&\leq \|x_0 - x^\star\|_{\mathbf{V}}^2
\end{aligned}
$$

위 부등식을 정리하면,

$$
L(x_T) - L(x^\star) \leq \frac{\alpha}{2}\left[\left(\frac{\gamma}{\gamma-\alpha}\right)^T - 1\right]^{-1}\|x_0 - x^\star\|_{\mathbf{V}}^2 .
$$

그러므로

$$
L(x_T) - L(x^\star) \leq \epsilon
$$

을 만족하는 T는 다음과 같다.

$$
T \geq \frac{\gamma}{\alpha}\log\left(\frac{\alpha}{2\epsilon}\|x_0 - x^\star\|_{\mathbf{V}}^2 + 1\right).
$$

□

6.2 $\mathbf{V}(x_t)$가 x_t에 의존할 때 MadProx의 전역 수렴성

Algorithm 2의 MadProx 갱신 x_t가 적절한 가정하에 최적해 $x^\star$로 전역 수렴함을 보일 것이다.

Theorem 6. *MadProx x_t가 아래의 가정들을 만족하면 임의의 $x_0 \in \mathrm{dom}\ L$을 시작점으로 설정했을 때 x_t는 $x^\star$로 수렴한다.*

1. *L이 닫힌 볼록 함수이고 $\inf_x\{L(x) \mid x \in \mathrm{dom}\ L\}$는 $x^\star$에서 존재한다.*

2. 모든 t에 대하여 $V(x_t) \succeq mI$을 만족하는 $m > 0$이 존재한다.

3. 모든 t에 대하여 $\Delta x_t = \mathrm{prox}_h^{\mathbf{V}(x_t)}\left(x_t - \mathbf{V}(x_t)^{-1}\nabla g(x_t)\right) - x_t$가 정확하게 계산된다.

Proof. 증명은 [3]의 Theorem 3.1을 참조하였다.

매 반복마다 $\eta \leq \min\left\{1, \frac{2m}{L_1}(1-\alpha)\right\}$를 만족하는 η_t를 선택하자. 그러면 Lemma 3에 의해 아래의 충분 하강 조건을 만족한다.

$$L(x_{t+1}) \leq L(x_t) + \alpha\eta_t\lambda_t,$$
$$\lambda_t := \nabla g(x_t)^T\Delta x_t + h(x_t + \Delta x_t) - h(x_t) \overset{(32)}{\leq} -\Delta x_t^T\mathbf{V}(x_t)\Delta x_t \leq 0,$$
$$\alpha \in (0, 0.5).$$

위 식으로부터 수열 $\{L(x_t)\}$은 증가하지 않는 수열임을 알 수 있다. $L(x)$는 닫힌 볼록 함수이고 최적해 $x^\star$가 존재하므로 비증가 수열 $\{L(x_t)\}$은 극한값을 가진다. 즉, $\eta_t\lambda_t$는 0으로 수렴한다. 그러면 모든 t에 대하여 $\eta_t > 0$이므로 λ_t가 0으로 수렴한다. 또한, 아래의 부등식으로부터 λ_t가 0으로 수렴하면 Δx_t 또한 0으로 수렴한다.

$$\|\Delta x_t\|^2 \leq \frac{1}{m}\Delta x_t^T\mathbf{V}(x_t)\Delta x_t \overset{(32)}{\leq} -\frac{1}{m}\lambda_t,$$
$$\lambda_t \to 0 \;\Rightarrow\; \Delta x_t \to 0.$$

그러므로 Proposition 4에 의해 x_t는 최적해 $x^\star$로 수렴함을 알 수 있다. □

6.3 $\mathbf{V}(x_t)$가 x_t에 의존할 때 지역 수렴성

이번에는 Algorithm 2의 MadProx 갱신 x_t가 적절한 조건하에서 반복 횟수 t가 충분히 클 때, 다음과 같이 q-quadratically 지역 수렴함을 보일 것이다:

$$\|x_{t+1} - x^\star\| = O\left(\|x_t - x^\star\|^2\right), \text{ for a sufficiently large } t.$$

Theorem 7. *모든 t에 대하여 $\mathbf{V}(x_t) \succeq mI$을 만족하는 상수 $m > 0$이 존재하고 $\mathbf{V}(x_t)$는 L_2-립시츠 연속이라고 하면, 다음의 전역 수렴 성질을 만족한다:*

$$\|x_{t+1} - x^\star\| = O(\|x_t - x^\star\|^2).$$

Proof. 증명은 [3]의 Theorem 3.4을 참조하였다.

η_t는 Algorithm 1의 직선 탐색 과정으로 구한 단계-크기라고 하자.

그러면 x_{t+1}은 다음과 같다:

$$x_{t+1} = x_t + \eta_t \Delta x_t = (1-\eta_t)x_t + \mathrm{prox}_h^{\mathbf{V}(x_t)}\left(x_t - \mathbf{V}(x_t)^{-1}\nabla g(x_t)\right).$$

Scaled proximal mapping의 firmly nonexpansive 성질에 의해 다음을 만족한다.

$$\begin{aligned}
\|x_{t+1} - x^\star\|_{\mathbf{V}(x_t)} &= \Big\|(1-\eta_t)x_t + \mathrm{prox}_h^{\mathbf{V}(x_t)}\left(x_t - \mathbf{V}(x_t)^{-1}\nabla g(x_t)\right) \\
&\quad - (1-\eta_t)x^\star - \mathrm{prox}_h^{\mathbf{V}(x_t)}\left(x^\star - \mathbf{V}(x_t)^{-1}\nabla g(x^\star)\right)\Big\|_{\mathbf{V}(x_t)} \\
&\le \Big\|\mathrm{prox}_h^{\mathbf{V}(x_t)}\left(x_t - \mathbf{V}(x_t)^{-1}\nabla g(x_t)\right) \\
&\quad - \mathrm{prox}_h^{\mathbf{V}(x_t)}\left(x^\star - \mathbf{V}(x_t)^{-1}\nabla g(x^\star)\right)\Big\|_{\mathbf{V}(x_t)} \\
&\le \left\|x_t - x^\star + \mathbf{V}(x_t)^{-1}(\nabla g(x^\star) - \nabla g(x_t))\right\|_{\mathbf{V}(x_t)} \\
&\le \frac{1}{\sqrt{m}}\left\|\mathbf{V}(x_t)(x_t - x^\star) - \nabla g(x_t) + \nabla g(x^\star)\right\|.
\end{aligned}$$

$\mathbf{V}(x_t)$는 L_2-립시츠 연속이므로,

$$\|\mathbf{V}(x_t)(x_t - x^\star) - \nabla g(x_t) + \nabla g(x^\star)\| \le \frac{L_2}{2}\|x_t - x^\star\|^2$$

이다. 따라서, x_t는 $x^\star$로 q-quadratically 수렴한다:

$$\|x_{t+1} - x^\star\| \le \frac{1}{\sqrt{m}}\|x_{t+1} - x^\star\|_{\mathbf{V}(x_t)} \le \frac{L_2}{2m}\|x_t - x^\star\|^2.$$

$$\Rightarrow \|x_{t+1} - x^\star\| = O(\|x_t - x^\star\|^2).$$

□

7 실험

7.1 실험 환경

본 논문의 실험은 R 3.6.3 버전을 사용하였으며, 이차 적률 행렬 $\mathbf{V}(w_t)$을 구할 때 expm v0.999-6 패키지의 sqrtm 함수를 이용해서 행렬 제곱근을 계산하였다. 실험의 계산은 Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz 단일 코어를 사용하였고, 188GB 메모리를 가지는 서버 환경에서 수행하였다.

표본 $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{0, 1\}$, $(i = 1, \cdots, n)$이 주어져 있을 때, $\ell 1$-노름 정규화 항이 있는 로지스틱 회귀 모형을 대상으로 세 가지 반복법(CD, MadProx, ProxNewton)의 수렴 속도를 비교하였다:

$$\underset{w \in \mathbb{R}^p}{\text{minimize}} \left[\frac{1}{n} \sum_{i=1}^{n} -\log f(y_i | \mathbf{x}_i, w) + \lambda \|w\|_1 \right],$$
$$\text{where } f(y_i|\mathbf{x}_i, w) = p_i^{y_i} \left(1 - p_i\right)^{1-y_i} \text{ and } p_i = \frac{\exp(\mathbf{x}_i^T w)}{1 + \exp(\mathbf{x}_i^T w)}.$$

참고문헌 [2]의 좌표 하강법(CD)을 기준(baseline)으로 하여 Algorithm 2의 MadProx, Algorithm 3의 ProxNewton의 수렴 속도를 정규분포에서 추출된 표본들을 이용한 시뮬레이션 환경에서 비교하였다. 시뮬레이션 과정은 [4]의 7. Numerical Experiments를 참고하였다. 시뮬레이션 데이터는 아래와 같이 총 네 가지 환경을 고려하였다.

1. 실험 1

   실험 1의 데이터 시뮬레이션 환경은 다음과 같다:

$$n = 500,\ p = 4,$$
$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\ I),\ w^\star \sim U(1,\ 10),$$
$$y_i \sim B(1,\ p_i^\star),\ p_i^\star = \frac{\exp(\mathbf{x}_i^T w^\star)}{1 + \exp(\mathbf{x}_i^T w^\star)}.$$

2. 실험 2

   실험 2의 데이터 시뮬레이션 환경은 다음과 같다:

$$n = 20000,\ p = 10,$$
$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\ I),\ w^\star \sim U(1,\ 10),$$
$$y_i \sim B(1,\ p_i^\star),\ p_i^\star = \frac{\exp(\mathbf{x}_i^T w^\star)}{1 + \exp(\mathbf{x}_i^T w^\star)}.$$

3. 실험 3

실험 3의 데이터 시뮬레이션 환경은 다음과 같다:

$$
\begin{aligned}
n = 500,\ p = 4,\\
\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\ \Sigma),\ w^\star \sim U(1,\ 10),\\
y_i \sim B(1,\ p_i^\star),\ p_i^\star = \frac{\exp(\mathbf{x}_i^T w^\star)}{1+\exp(\mathbf{x}_i^T w^\star)}.
\end{aligned}
$$

여기서 Σ의 조건수는 100.0203이다.

4. 실험 4

실험 4의 데이터 시뮬레이션 환경은 다음과 같다:

$$
\begin{aligned}
n = 20000,\ p = 10,\\
\mathbf{x}_i \sim \mathcal{N}(\mathbf{0},\ \Sigma),\ w^\star \sim U(1,\ 10),\\
y_i \sim B(1,\ p_i^\star),\ p_i^\star = \frac{\exp(\mathbf{x}_i^T w^\star)}{1+\exp(\mathbf{x}_i^T w^\star)}.
\end{aligned}
$$

여기서 Σ의 조건수는 100.0809이다.

위 실험 환경을 정리하면 다음과 같다. 실험 1의 표본은 공분산 행렬이 단위행렬인 정규분포에서 추출되었고 표본의 수(n), 모수의 차원(p)이 모두 작다. 실험 2의 표본은 공분산 행렬이 단위행렬인 정규분포에서 추출되었고 표본의 수(n), 모수의 차원(p)이 모두 크다. 실험 3의 표본은 조건수가 약 100인 공분산 행렬을 가지는 정규분포에서 추출되었고 표본의 수(n), 모수의 차원(p)이 모두 작다. 실험 4의 표본은 조건수가 약 100인 공분산 행렬을 가지는 정규분포에서 추출되었고 표본의 수(n), 모수의 차원(p)이 모두 크다.

7.2 실험 결과

실험 결과에 대한 그림에는 각 반복법의 반복 횟수 t의 로그값 $\log_{10} t$을 x축으로, 좌표하강법으로 구한 모수 추정값 $w_{\ell 1}$과 각 반복법 w_t의 차이에 대한 노름에 로그를 씌운 값인 $\log_{10} \|w_t - w_{\ell 1}\|_2$와 손실함수 값의 차이에 로그를 씌운 값인 $\log_{10} L(w_t) - L(w_{\ell 1})$을 y축으로 하여 세 반복법 CD, ProxNewton, MadProx의 수렴 속도를 비교하였다.

위 4개 실험의 결과에 대한 그림에서 세 반복법 CD, ProxNewton, MadProx 모두 오차(y축)가 매우 작은 값을 가지는 것으로 보았을 때 MadProx 갱신 w_t은 최적값 $w_{\ell 1}$으로 수렴한다는 것을 알 수 있다. 즉, 실험 환경에서 MadProx 갱신은 Theorem 6의 전역 수렴성을 만족한다.

또한, 네 가지 실험 환경에서 MadProx 갱신의 수렴 속도는 모두 ProxNewton과 좌표하강법(CD)의 사이에 있다. 다시 말해서, 표본의 수(n)와 모수의 차원(p)이 작을 때(Figure

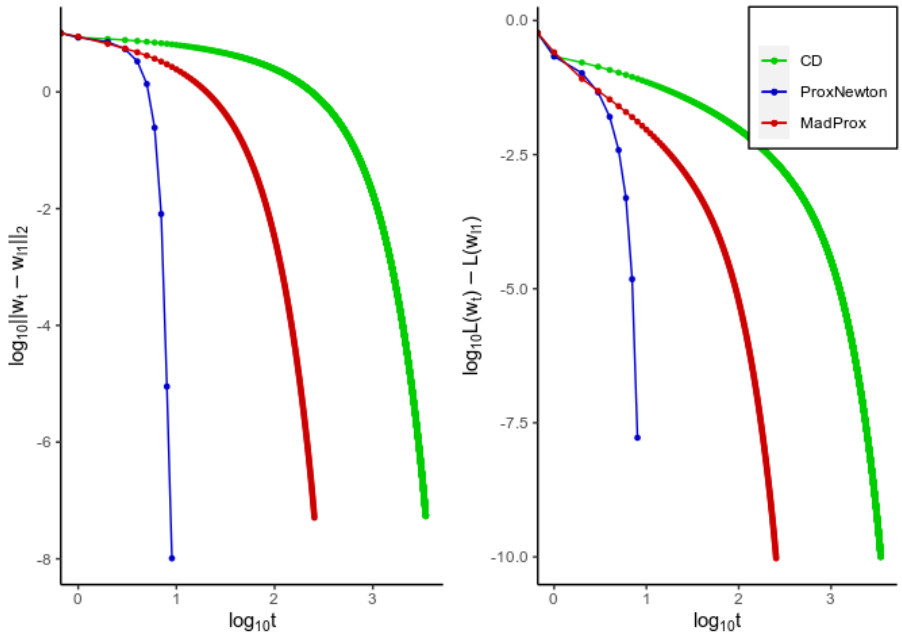

Figure 1: Experiment 1 ($\lambda = 0.001$)

1, Figure 3) MadProx 갱신의 수렴 속도는 ProxNewton과 좌표 하강법(CD)의 사이에 있다. 표본의 수(n)와 모수의 차원(p)이 클 때(Figure 2, Figure 4) MadProx 갱신의 수렴 속도는 ProxNewton과 좌표 하강법(CD)의 사이에 있다. 그리고 공분산 행렬 Σ의 조건수가 클 때 (Figure 3, Figure 4)에도 MadProx 갱신의 수렴 속도는 ProxNewton과 좌표 하강법(CD)의 사이에 있다. 즉, 1차 수렴 속도를 가진 것으로 알려진 좌표 하강법(CD)보다 수렴 속도가 빠르고 2차 수렴 속도를 가진 것으로 알려진 헤시안 행렬을 이용한 ProxNewton보다는 수렴 속도가 느리다. 이로서 실험 환경에서 MadProx 갱신은 Theorem 7의 q-quadratically 지역 수렴성을 어느정도 만족하는 것을 알 수 있다.

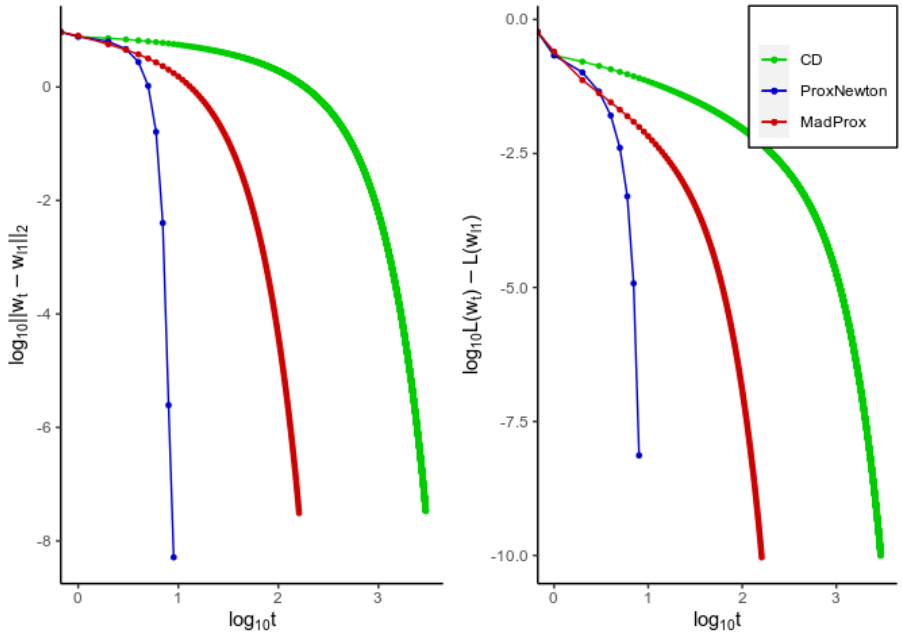

Figure 2: Experiment 2 ($\lambda = 0.001$)

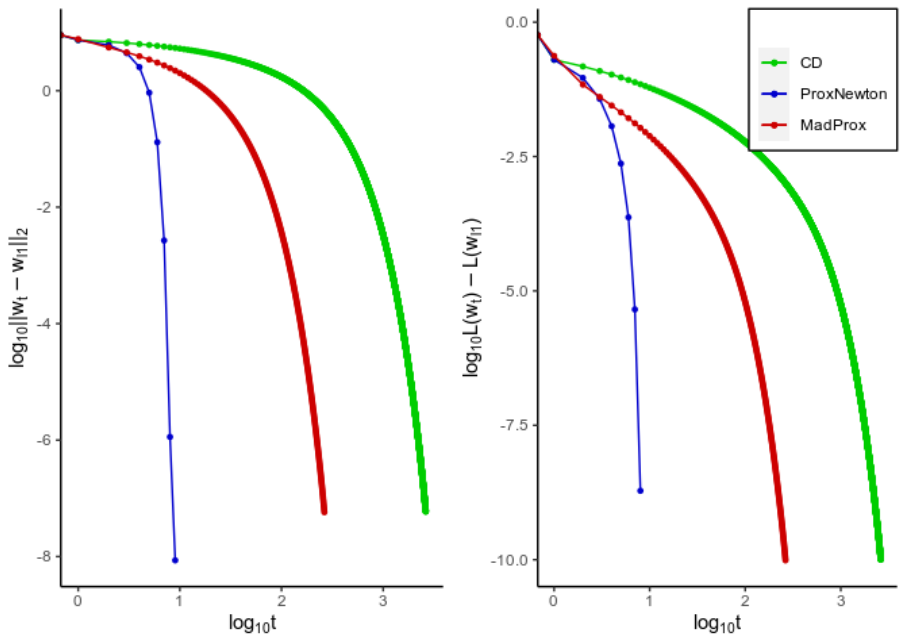

Figure 3: Experiment 3 ($\lambda = 0.001$)

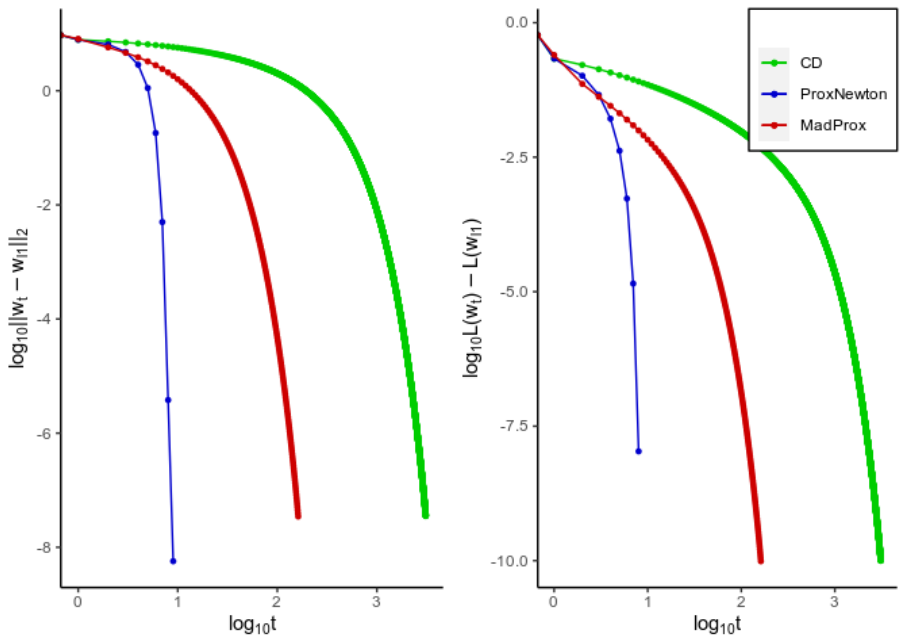

Figure 4: Experiment 4 ($\lambda = 0.001$)

8 결론

MasGrad는 이차 정보로 이차 적률 행렬을 사용하는 이차 반복법이다. MasGrad는 다른 이차 반복법인 뉴턴 방법과 같이 특정 조건 하에서 네스테로프 가속 성질을 가지고 있고 MasGrad 갱신의 결합 분포와 가까운 갱신인 diff_MasGrad 갱신을 이용하여 통계적 추론을 할 수 있다. 특히, 매끄럽지 않은 정규화 항이 있는 목적함수를 최적화하는 상황에서, 적률 조정 근위 함수를 사용하여 MasGrad 갱신과 유사한 반복법인 MadProx를 정의하였다. 참고문헌 [4]의 MadProx는 제한된 상황, 즉 $\mathbf{V}(x_t) = \mathbf{V}$일 때의 수렴성을 보였지만 본 논문에서는 $\mathbf{V}(x_t)$가 x_t에 의존할 때의 수렴성을 실험을 통해 보였다. 다시 말해서, 총 4개의 인위적인 데이터를 만들어 $\ell 1$-노름 정규화 항이 있는 로지스틱 회귀 모형을 대상으로 MadProx의 수렴 여부와 수렴 속도를 보였다. 한편, MadProx는 매 반복마다 이차 적률 행렬과 scaled proximal mapping을 계산하기 때문에 계산 복잡도가 큰 편이다. Quasi-Newton 방법과 같이 이차 적률 행렬의 근삿값을 구하는 방법이 제안된다면, 속도와 안정성 측면에서 어떤 방법이 더 효율적인지 검증해볼 필요가 있다.

9 부록

Non-asymptotic bound for inference

MasGrad process(θ_t)의 결합 분포를 $\mu(\theta_t, t \in [T])$, diff_MasGrad process(ξ_t)의 결합 분포를 $\mu(\xi_t, t \in [T])$라고 하자. Classic CLT로는 두 확률 과정의 분포에 대한 KL divergence(relative entropy)를 구할 수 없기에 entropic CLT를 이용한다. entropic CLT를 이용하기 위해 아래의 두 가정이 필요하다.

(1) 정규 분포에 대해 절대 연속성: 어떤 상수 D_1에 대해 랜덤 벡터 $X \in \mathbb{R}^p$ 가 정규 분포와 entropic distance에 대해 유계라고 가정하자.

$$D_{\mathrm{KL}}(\mu(X) \| \mu(\mathbf{g})) < D_1, \text{ where } \mathbf{g} \sim \mathcal{N}(0, I_p).$$

(2) 유한한 $(4+\delta)$차 모멘트: 다음을 만족하는 상수 D_2가 존재한다고 가정하자.

$$\mathbb{E} \|X\|^{4+\delta} < D_2, \text{ for some small } \delta > 0.$$

참고문헌

[1] Sergey G Bobkov, Gennadiy P Chistyakov, and Friedrich Götze. "Rate of convergence and Edgeworth-type expansion in the entropic central limit theorem". In: *The Annals of Probability* (2013), pp. 2479–2512.

[2] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent". In: *Journal of statistical software* 33.1 (2010), p. 1.

[3] Jason D Lee, Yuekai Sun, and Michael A Saunders. "Proximal Newton-type methods for minimizing composite functions". In: *SIAM Journal on Optimization* 24.3 (2014), pp. 1420–1443.

[4] Tengyuan Liang and Weijie J Su. "Statistical inference for the population landscape via moment-adjusted stochastic gradients". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.2 (2019), pp. 431–456.

[5] Alexandre B Tsybakov. *Introduction to Nonparametric Estimation.* 2009.

Abstract

Ahn Woo Jin
Department of Statistics
The Graduate School
Seoul National University

In this paper, we introduce MadProx method that optimizes the objective function with a non-smooth normalization term and analyze its convergence. Prior to MadProx, we will introduce MasGrad method, which is the basis of MadProx, and prove the statistical inference properties of MasGrad. MadProx method has a limitation in that it converges under limited conditions, such as when the second moment matrix does not depend on the updated variable. In this paper, we will define a new search direction and introduce an algorithm that chooses a step-size at each iteration through a line search procedure rather than a fixed step-size. We will show MadProx with the search direction and step-size obtained through the line search procedure converges even if the second moment matrix depends on the updated variable. In addition, the convergence rate of MadProx will be compared with other iterative methods using logistic regression with L1-regularization terms as a representative model in a data simulation environment.

keywords: optimization, iterative method, MasGrad, MadProx, proximal function, line search procedure
Student Number: 2020-21707