ㅏ 댐: 오전 ETHOD FOR 에스 TOCHASTIC 영형 최적화

Diederik P. Kingma *
암스테르담 대학교 OpenAI
dpkingma@openai.com

지미 레이 바 *
토론토 대학교
jimmy@psi.utoronto.ca

ㅏ BSTRACT

소개합니다 *아담,* 저차 모멘트의 적응 적 추정에 기반한 확률 적 목적 함수의 1 차 기울기 기반 최적화를위한 알고리즘. 이 방법은 구현이 간단하고 계산 효율이 높으며 메모리 요구 사항이 적고 기울기의 대각선 크기 조정에 불변하며 데이터 및 / 또는 매개 변수 측면에서 큰 문제에 적합합니다. 이 방법은 또한 비정상적인 대물 렌즈 및 매우 시끄러운 및 / 또는 희소 기울기가있는 문제에 적합합니다. 하이퍼 파라미터는 직관적 인 해석이 가능하며 일반적으로 약간의 조정이 필요합니다. 관련 알고리즘에 대한 일부 연결, *아담* 영감을 얻었습니다. 또한 알고리즘의 이론적 수렴 속성을 분석하고 온라인 볼록 최적화 프레임 워크에서 가장 잘 알려진 결과에 필적하는 수렴 률에 대한 후회를 제공합니다. 경험적 결과는 Adam이 실제로 잘 작동하고 다른 확률 적 최적화 방법과 유리하게 비교됨을 보여줍니다. 마지막으로 *AdaMax,* 의 변형 *아담* 인피니티 규범을 기반으로합니다.

1 나는 소개

확률 적 기울기 기반 최적화는 많은 과학 및 공학 분야에서 실질적으로 핵심적인 중요합니다. 이러한 필드의 많은 문제는 매개 변수에 대해 최대화 또는 최소화가 필요한 일부 스칼라 매개 변수화 된 목적 함수의 최적화로 캐스팅 될 수 있습니다. 함수가 매개 변수에 따라 미분 할 수있는 경우 경사 하강 법은 상대적으로 효율적인 최적화 방법입니다. 모든 매개 변수에 대한 1 차 편도 함수의 계산은 함수를 평가하는 것과 동일한 계산 복잡도를 갖기 때문입니다. 종종 목적 함수는 확률 적입니다. 예를 들어, 많은 목적 함수는 서로 다른 데이터 하위 샘플에서 평가 된 하위 함수의 합계로 구성됩니다. 이 경우 기울기 단계를 수행하여 최적화를보다 효율적으로 만들 수 있습니다.

wrt 개별 하위 기능, 즉 확률 적 경사 하강 법 (SGD) 또는 상승. SGD는 최근 딥 러닝의 발전과 같은 많은 머신 러닝 성공 사례에서 중심이되는 효율적이고 효과적인 최적화 방법임을 입증했습니다 (Deng et al., 2013; Krizhevsky et al., 2012; Hinton & Salakhutdinov, 2006; Hinton et al., 2013; al., 2012a; Graves et al., 2013). 목표에는 드롭 아웃 (Hinton et al., 2012b) 정규화와 같은 데이터 서브 샘플링 이외의 다른 노이즈 소스도있을 수 있습니다. 이러한 모든 잡음이있는 목표에 대해 효율적인 확률 적 최적화 기술이 필요합니다. 이 백서의 초점은 고차원 매개 변수 공간을 사용한 확률 적 목표의 최적화에 있습니다. 이러한 경우 고차 최적화 방법은 부적합하며이 백서에서의 논의는 1 차 방법으로 제한됩니다.

우리는 제안한다 *아담,* 메모리 요구 사항이 거의없는 1 차 그래디언트 만 필요로하는 효율적인 확률 적 최적화 방법. 이 방법은 기울기의 첫 번째 모멘트와 두 번째 모멘트의 추정에서 다른 매개 변수에 대한 개별 적응 학습률을 계산합니다. 이름 *아담*
적응 모멘트 추정에서 파생됩니다. 우리의 방법은 최근에 널리 사용되는 두 가지 방법의 장점을 결합하도록 설계되었습니다. AdaGrad (Duchi et al., 2011)는 희소 그래디언트와 잘 작동하고 RMSProp (Tieleman & Hinton, 2012)는 온라인에서 잘 작동합니다. 및 비 고정 설정; 이들 및 기타 확률 적 최적화 방법에 대한 중요한 연결은 섹션 5에 설명되어 있습니다. Adam의 장점 중 일부는 매개 변수 업데이트의 크기가 기울기의 크기 조정에 불변하고, 단계적 크기가 단계적 하이퍼 파라미터에 의해 거의 제한되며, 고정이 필요하지 않다는 것입니다. 객관적으로 희소 그라디언트로 작동하며 자연스럽게 단계 크기 어닐링의 형태를 수행합니다.

* 동등한 기여. 저자 순서는 Google 행 아웃을 통한 코인 fl ip에 의해 결정됩니다.

알고리즘 1 : *아담,* 확률 적 최적화를 위해 제안 된 알고리즘입니다. 자세한 내용은 섹션 2를 참조하십시오.
그리고 약간 더 효율적인 (그러나 덜 명확한) 계산 순서를 위해. $지/2$ $티/$요소 별을 나타냅니다.
광장 $지/티/$ $지/티/$ 테스트 된 기계 학습 문제에 대한 좋은 기본 설정은 다음과 같습니다. $\alpha = 0.001$,
$\beta_1 = 0.9$, $\beta_2 = 0.999$ 과 $\varepsilon = 10 - 8.$ 벡터에 대한 모든 연산은 요소 단위입니다. 와 β $티/$ 1과 β $티/$ 2
우리는 β_1 과 β_2 권력에 *티/.*

필요 : α : 스텝 사이즈
필요 : $\beta_1, \beta_2 \in [0, 1)$: 순간 추정치에 대한 지수 감쇠율
필요 : $f(\theta)$: 매개 변수가있는 확률 적 목적 함수 θ
필요 : θ_0 : 초기 매개 변수 벡터
$미/디/업_0 \leftarrow 0$ (초기화 1 성 모멘트 벡터)
$V_0 \leftarrow 0$ (2 초기화 nd 모멘트 벡터)
$티/ \leftarrow 0$ (타임 스텝 초기화)
동안 θ $티/$수렴되지 않음 하다
$티/ \leftarrow t + 1$
$지/티/ \leftarrow \nabla_\theta$ *에프* $티/(\theta$ $티/-1)$ (시간 단계에서 확률 적 목표에 대한 그라디언트 가져 오기 *t)* m $티/ \leftarrow \beta_1 \cdot 미/디/업$ $티/$
$-1 + (1 - \beta_1) \cdot 지/t$ (편향된 첫 번째 순간 추정 업데이트)
V $티/ \leftarrow \beta_2 \cdot V$ $티/-1 + (1 - \beta_2) \cdot 지/2$ (편향된 둘 번째 원시 모멘트 추정 업데이트)
$미/디/업$ $티/ \leftarrow 미/디/업$ $티/(1 - \beta_1$ *t)* (편향 보정 된 첫 번째 순간 추정값 계산)
V $티/ \leftarrow V$ $티/(1 - \beta_2$ *t)* (공동 $\sqrt{}$ mpute bias-corrected second raw moment 추정)
θ $티/ \leftarrow \theta$ $티/-1 - \alpha \cdot 미/디/업$ $티/(\sqrt{V_t} + \varepsilon)$ (매개 변수 업데이트)
동안 끝
반환 θ_t (결과 매개 변수)

섹션 2에서는 알고리즘과 업데이트 규칙의 속성에 대해 설명합니다. 섹션 3은 초기화 바이어스 보정 기술을 설명하고 섹션 4는 온라인 볼록 프로그래밍에서 Adam의 수렴에 대한 이론적 분석을 제공합니다. 경험적으로 우리의 방법은 섹션 6에 표시된 것처럼 다양한 모델 및 데이터 세트에 대해 다른 방법보다 지속적으로 성능이 뛰어납니다. 전반적으로 Adam이 대규모 고차원 기계 학습 문제로 확장되는 다목적 알고리즘임을 보여줍니다.

2A LGORITHM

제안 된 알고리즘의 의사 코드는 알고리즘 1 참조 *아담.* 허락하다 $f(\theta)$ 잡음이있는 목적 함수 : 미분 할 수있는 wrt 매개 변수 인 확률 적 스칼라 함수 θ. 우리는이 함수의 기대 값을 최소화하는 데 관심이 있습니다. 이자형[$f(\theta)$] wrt 매개 변수 θ. 와

에프 $1(\theta)$, ..., $f_T(\theta)$ 우리는 후속 시간 단계에서 확률 함수의 실현을 나타냅니다.
1, ..., *T.* 확률 성은 무작위 서브 샘플 (미니 배치)의 평가에서 올 수 있습니다.
또는 고유 한 기능 잡음에서 발생합니다. 와 $지/t = \nabla_\theta$ *에프* $티/(\theta)$ 우리는 기울기, 즉 편도 함수의 벡터를 나타냅니다. *에프* $티/$ wrt θ
시간 단계에서 평가 *티/.*

알고리즘은 기울기의 지수 이동 평균을 업데이트합니다 ($미/디/업$ $티/$) 제곱 그래디언트
(V $티/$) 하이퍼 파라미터는 $\beta_1, \beta_2 \in [0, 1)$ 이러한 이동 평균의 지수 감쇠율을 제어합니다. 이동 평균 자체는 1의 추정치입니다. 성
순간 (평균)과
2 nd 그래디언트의 원시 모멘트 (비 중심 분산). 그러나 이러한 이동 평균은 0의 (벡터)로 초기화되어 특히 초기 시간 단계
동안, 특히 감쇠율이 작은 경우 (즉, βs는 1)에 가깝습니다. 좋은 소식은이 초기화 편향이 쉽게 상쇄 될 수 있다는 것입니다.

견적 $미/디/업$ $티/$과 V $티/$ 자세한 내용은 섹션 3을 참조하십시오.

알고리즘 1의 효율성은 명확성을 희생하면서 다음을 변경함으로써 개선 될 수 있습니다.
계산 순서 (예 : 마지막 세 개를 대체) $\sqrt{}$다음 줄이있는 루프의 줄 :
$\alpha_t = \alpha \cdot \sqrt{1 - \beta_2}$ $티/$ $2 / (1 - \beta_1$ $티/$ $1)$과 θ $티/ \leftarrow \theta$ $티/-1 - \alpha$ $티/ \cdot 미/디/업$ $티/(\sqrt{V_t} + \hat{\varepsilon})$.

2.1A 댐 ' S 업데이트 규칙

Adam의 업데이트 규칙의 중요한 속성은 신중한 선택입니다. $\sqrt{}$ 계단식. 가정 $\varepsilon = 0$, 그만큼
매개 변수 공간에서 취한 효과적인 단계 $\sqrt{}$ 타임 스텝 $티/$이다 $\Delta_t = \alpha \cdot 미/디/업$ $티/$ V $티/$ 효과적인 Stepize는
두 개의 상한 : $|\Delta_t| \leq \alpha \cdot (1 - \beta_1)/\sqrt{1 - \beta_2}$ 경우에 $(1 - \beta_1) > \sqrt{1 - \beta_2}$ $\sqrt{1 - \beta_2}$ 및 $|\Delta_t| \leq \alpha$

그렇지 않으면. 첫 번째 경우는 희소성이 가장 심한 경우에만 발생합니다. 그래디언트가 th를 제외한 모든 시간 단계에서 0
인 경우 √
더 작아 질 것입니다. 언제 (1 − $\beta_{1)}$ = 1 − β_2 w √ e 현재 시간 단계. 덜 드문 경우의 경우 효과적인 단계 화 전자가 | *미디/엄* $_{E//}$ $V_{t/}$ < 1 따라서 | $\Delta_{t/}$ < *α.* 에
더 일반적인 시나리오, 우리는 *미디/엄* $_{E//}$ $V_{E/}$ ≈ ± 1 이후 | 이자형[*지*]/ 이자형[*지*/$_{2|}$ ≤ 1. 각 시간 단계에서 매개 변수 공간에서
취한 단계의 유효 크기는 대략 다음과 같이 제한됩니다.
Stepize 설정 *α,* 즉, | $\Delta_{t/}$ / *α.* 이것은 설정으로 이해할 수 있습니다 *신뢰/ 지/역* 현재의 기울기 추정치가 충분하지 않은 현재
매개 변수 값 주변
정보. 이것은 일반적으로 적절한 척도를 비교적 쉽게 알 수 있도록합니다. *α* 미리. 예를 들어, 많은 기계 학습 모델에서
우리는 종종 좋은 최적화가 매개 변수 공간의 일부 설정 영역 내에서 높은 확률을 갖는다는 것을 미리 알고 있습니다. 예를
들어, 모수에 대한 사전 분포를 갖는 것은 드문 일이 아닙니다. 이후 *α* 매개 변수 공간에서 단계의 크기를 설정 (상한)하면
종종 올바른 크기의 순서를 추론 할 수 있습니다. *α* 그런 최적화

에서 도달 할 수 있습니다 θ_0 w √ 몇 번의 반복이 있습니다. 용어를 약간 남용하면 비율을 *미디/엄* $_{E//}$ $V_{E/}$ 그만큼 *신호 대 잡음* 비율
(*SNR).* SNR이 작을수록 효과적입니다.
단계적으로 $\Delta_{E/}$ 0에 가까워집니다. 더 작은 SNR은 다음을 의미하기 때문에 이것은 바람직한 속성입니다.
방향 여부에 대한 더 큰 불확실성이 있습니다. *미디/엄* $_{E/}$ 실제 그라디언트의 방향에 해당합니다. 예를 들어, SNR 값은
일반적으로 최적을 향해 0에 가까워지고
매개 변수 공간에서 더 작은 유효 단계 : 자동 어닐링의 한 형태. 효과적인 Stepize
$\Delta_{E/}$ 또한 그래디언트의 규모에도 변하지 않습니다. 그 레이디 크기 조정 √ nts *지* 요인으로 *씨*/ 확장됩니다 *미디/엄* $_{E/}$
요인으로 *씨*/과 $V_{E/}$ 요인으로 *씨*/$_{2,}$ 취소 : (*씨*/ · *미디/엄* $_{E//}$)/(*씨*/$_2$ · $V_{t)}$ = *미디/엄* $_{E//}$ $V_{E/.}$

3 I NITIALIZATION BIAS CORRECTION

섹션 2에서 설명한대로 Adam은 초기화 바이어스 보정 용어를 사용합니다. 여기서 우리는 2 차 모멘트 추정에 대한 용어를
유도 할 것입니다. 첫 번째 순간 추정에 대한 유도는 완전히 유사합니다. 허락하다 *지*/ 확률 적 목표의 기울기 *에프,* 제곱
그래디언트의 지수 이동 평균을 사용하여 두 번째 원시 모멘트 (비 중심 분산)를 추정하려고합니다.

부패율로 $\beta_{2.}$ 허락하다 *지*/$_{1, ...,}$ *지*/$_{E/}$ 후속 타임 스텝의 그래디언트이고, 각각은 기본 그래디언트 분포에서 가져온 것입니다. *지*/$_{E/}$
~ *p (g* $_{E/),}$ 지수 이동 평균을 다음과 같이 초기화하겠습니다.
V_0 = 0 (0으로 구성된 벡터). 먼저 타임 스텝의 업데이트가 *E*/ 지수 이동 평균의
V_t = β_2 · $V_{E/}$ − 1 + (1 − $\beta_{2)}$ · *지*/$_2$ $_{t\,(}$ 어디 *지*/$_2$ $_{E/}$ 요소 별 사각형을 나타냅니다. *지*/ $_{E/}$ *지*/$_{E/)}$ 다음과 같이 쓸 수 있습니다.
이전의 모든 시간 단계에서 그라디언트의 함수 ∑ :

$$V_{t\,(} = (1 - \beta_2) \sum_{i=1}^{E/} \beta_{2}^{E/ - \text{나는}} \cdot \text{지}^{2}_{\text{나는}} \quad (1)$$

우리는 방법을 알고 싶습니다 이자형[$V_{E/),}$ 시간 단계에서 지수 이동 평균의 예상 값 *E/,*
진정한 두 번째 순간과 관련이 있습니다. 이자형[*지*/$_2$ $_{t),}$ 둘 사이의 불일치를 수정할 수 있습니다.
왼손과 오른손에 대한 기대 ∑ - eq의 ides. (1]):

$$\text{이자형}[V_{t)} = \text{E}\,(1 - \beta_2) \sum_{i=1}^{E/} \beta_{2}^{\text{나는}} \cdot \text{지}_{2\ \text{나는}} \quad (2)$$

$$= \text{이자형}[\text{지}^{2}_{2}](1 - \beta_2) \sum_{i=1}^{E/} \beta_{2}^{E/ - \text{나는}} + \zeta \quad \text{(삼)}$$

$$= \text{이자형}[\text{지}^{2}_{2}](1 - \beta_{E/\ 2)} + \zeta \quad (4)$$

어디 ζ = 0 진정한 두 번째 순간이 이자형[*지*/$_2$ $_{//}$ 고정되어 있습니다. 그렇지 않으면 ζ 작게 유지할 수 있습니다.
지수 붕괴율 β_1 지수 이동 평균이 과거에 너무 먼 기울기에 작은 가중치를 할당하도록 선택할 수 있습니다 (그리고
선택해야합니다). 남은 것은 (1 − $\beta_{E/}$ $_{2)}$ 그것은
실행 평균을 0으로 초기화하여 발생합니다. 따라서 알고리즘 1에서는 초기화 편향을 수정하기 위해이 항으로 나눕니다.

희소 기울기의 경우, 두 번째 모멘트의 신뢰할 수있는 추정을 위해 평균을
작은 값을 선택하여 많은 그라디언트 $\beta_{2;}$ 그러나 그것은 정확히이 작은 경우입니다 β_2 초기화 편향 보정이 없으면 훨씬 더
큰 초기 단계로 이어집니다.

4C 온 버전 스 분석

제안 된 온라인 학습 프레임 워크를 사용하여 Adam의 융합을 분석합니다 (Zinkevich, 2003). 임의의 알려지지 않은 볼록 비용 함수 시퀀스가 주어짐 $에프_1(\theta), f_2(\theta), ..., f_T(\theta)$. 매번 $티$, 우리의 목표는 매개 변수를 예측하는 것입니다. $\theta_티$ 이전에 알려지지 않은 비용 함수에서 평가합니다. $에프_티$. 시퀀스의 특성을 미리 알 수 없기 때문에 알고리즘을 평가합니다.

후회를 사용하여 이전의 모든 차이의 합계입니다. 엑스 e 온라인 예측 사이 $에프_티(\theta_티)$ 최적의 고정 점 매개 변수 $에프_티(\theta^*)$ 이리저리 ∑ 이전의 모든 단계에 대해 가능한 설정입니다. 구체적으로 후회는 다음과 같이 정의됩니다.

$$R(T) = \sum_{t=1}^{티} [에프_티(\theta_티) - 에프_티(\theta^*)] \quad (5)$$

어디 $\theta^* = 아르\ 민_{\theta \in X}$ ∑ $티$ √ —

부록에서. 우리의 결과 $_{t=1}$ $에프_티(\theta)$이 일반적인 (볼) 록 원라 있 종매해 취 행 절 알려진 경계와 비슷합니다.

학습 문제. 또한 몇 가지 정의를 사용하여 표기법을 단순화합니다. $지_{t,}$ ∇ $에프_티(\theta_티)$과 $지_{t,i}$ 로 $나는$일 요소. 우리는 정의 $지_{1:t,i} \in$ 아르 자형 $_{티}$포함하는 벡터로 $나는$일 그라디언트의 차원 모든 반복에 걸쳐 $t, g_{1:t,i} = [지_{1,나는}, 지_{2,나는}, \cdots, 지_{t,i}]$. 또한 우리는 $\gamma, \sqrt{\beta_1}$. 우리의 다음 $\frac{2}{\beta_2}$ 정리는 학습률이 $\alpha_{티}$ 속도로 쇠퇴하고 있습니다 $티^{-1}$ ^{2}그리고 첫 순간 실행 평균 계수 $\beta_{1,티}$ 기하 급수적으로 쇠퇴하다 λ, 일반적으로 1에 가깝습니다. 예 : $1 - 10^{-8}$.

정리 4.1.

$지_\infty$ 모든 $\theta \in$ 함수가 $에프$ 최적 계 경우에는 그 때문에 다음과 같습니다. 그래서 $에프$에 의해 생성 된, 매개 변수 $\theta_2 \leq$ |,

$\|\theta_{미디엄} - \theta_{엔}\|_\infty \leq D_\infty$ 어떠한 것도 $m, n \in \{1, ..., 티\}$, 과 $\beta_1, \beta_2 \in [0, 1)$ 폴다 $\sqrt{\beta_2}$ $\frac{1 \leq}{\beta_2}$1. 허락하다 $\alpha_t = \sqrt{}$ $\frac{\alpha}{티}$ 과 $\beta = \beta_{1,}$ $_{티-1}$, $\lambda_1 \in (0, 1)$. Adam은 모두에게 다음과 같은 보증을 제공합니다. $티 \geq 1$.

$$R(T) \leq \frac{D_2}{2\alpha(1\beta)_1} \sum_{i=1}^{D\sqrt{}} \overline{TV_{T,i}} - \beta_1 1 - \beta(1-\gamma)_2 \frac{\alpha(1\sqrt{}+\beta)G_\infty}{(1} \sum^{D} \|지_{1:T,i}\|_2 + \|\sum^{D} D_2 \, 지_{\infty} 1 - \beta \frac{\sqrt{}}{2\alpha(1-\beta_1)(1-_2 \lambda)_2}$$

우리 T √heof ême 4.1은 데이터 특성이 희소하고 경계가있는 기울기 인 경우를 의미합니다. √ 그는 합계- 결합 용어는 mu 일 수 있습니다. √ ch가 상한보다 작음 $\sum_{i=1}^{D} \overline{TV_{T,i}} \ll dG_\infty T$, 특히 함수와 데이터의 클래스가 ∑ 기능은 다음과 같은 형태입니다.

섹션 1.2 (Duchi et al., 2011). 기대 값에 대한 결과 이차형[$_{D}$ $\sum_{i=1}^{} \|지_{1:T,i}\|_2$] 또한 √플리 아담에게. 특히, th √ Adam 및 Adagrad와 같은 적응 형 방법은 $영형(\log dT)$, 개선 $O(dT)$ 비 적응 방법의 경우. 부패 $\beta_{1,티}$0으로 향하는 것은 우리의 이론적 분석에서 중요하며 이전의 경험적 결과와도 일치합니다. 예를 들어 (Sutskever et al., 2013)은 훈련이 끝날 때 모멘텀 계수를 줄이면 수렴을 향상시킬 수 있다고 제안합니다. 마지막으로 아담의 평균적인 후회가 수렴되는 것을 보여줄 수 있습니다.

추론 4.2. t 기능 $에프_티$ 경계가있는 그라디언트가 있습니다. $\|\nabla 에프_티(\theta)\|_2 \leq 지, \|\nabla 에프_티(\theta)\|_\infty \leq$ 또는 al $_{D}$ $_{\infty}$ 그쪽으로 ssume 그리고 거리 사이 ee {n any $\theta_티$} Adam에 의해 생성 된 $\|\theta_{엔} - \theta_{미디엄}\|_2 \leq D, D_\infty$ 어떠한 것도 m, n " 회대에프 − " 엘 θ ≤ 아르 자형 1, ..., T. Adam은 모두에게 다음과 같은 보증을 제공합니다. $티 \geq \theta$ 엔 ∞

$$\frac{R(T)}{티} = 영형\left(\frac{1}{\sqrt{}}\right) + \frac{1}{티 \sum D}$$

이 결과는 Theorem 4.1을 사용하여 얻을 수 있습니다. √ — $\sum_{i=1} \|지_{1:T,i}\|_2 \leq dG_\infty T$. 그러므로, 임 $_{\to 티} R(T) = 0$. $_{티}$

5R 고양 된 일

Adam과 직접적인 관계가있는 최적화 방법은 RMSProp (Tieleman & Hinton, 2012; Graves, 2013) 및 AdaGrad (Duchi et al., 2011)입니다. 이러한 관계는 아래에서 설명합니다. 다른 확률 적 최적화 방법에는 vSGD (Schaul et al., 2012), AdaDelta (Zeiler, 2012) 및 Roux & Fitzgibbon (2010)의 natural Newton 방법이 포함되며 모든 설정은 곡률을 추정하여 단계 화됩니다.

첫 번째 주문 정보에서. SFO (Sum-of-Functions Optimizer) (Sohl-Dickstein et al., 2014)는 미니 배치를 기반으로하는 준 뉴턴 방법이지만 (Adam과 달리) 데이터 세트의 미니 배치 파티션 수에 선형 메모리 요구 사항이 있습니다. GPU와 같이 메모리가 제한된 시스템에서는 종종 실행 불가능합니다. 자연 경사 하강 법 (NGD) (Amari, 1998)과 마찬가지로 Adam은

데이터의 기하학에 적응하는 전제 조건 $V_{E/}$Fisher 정보 매트릭스의 대각선에 대한 근사치입니다 (Pascanu & Bengio, 2013). 그러나 Adam의 전제 조건 (예 :
AdaGrad 's)는 대각선 Fisher 정보 행렬 근사값의 역의 제곱근으로 사전 조정하여 바닐라 NGD보다 적응에서 더 보수적입니다.

RMSProp : Adam과 밀접하게 관련된 최적화 방법은 RMSProp (Tieleman & Hinton,
2012). 모멘텀이있는 버전이 때때로 사용되었습니다 (Graves, 2013). 모멘텀이있는 RMSProp와 Adam 사이에는 몇 가지 중요한 차이점이 있습니다. 모멘텀이있는 RMSProp는 재조정 된 기울기의 모멘텀을 사용하여 매개 변수 업데이트를 생성하는 반면 Adam 업데이트는 기울기의 첫 번째 및 두 번째 모멘트의 실행 평균을 사용하여 직접 추정됩니다. .
RMSProp
또한 편향 보정 용어가 없습니다. 이것은 값의 경우에 가장 중요합니다 β_2 1에 가까움 (희소 기울기의 경우 필요),이 경우 바이어스를 수정하지 않으면 매우 큰 스텝 화가 발생하고
섹션 6.4에서 경험적으로 입증했듯이 종종 발산합니다.

AdaGrad : 희소 GR에 잘 작동하는 알고리즘 $\sqrt{\sum}$ $_{E/}$ adients는 AdaGrad입니다 (Duchi et al., 2011). 이것의

기본 버전은 매개 변수를 다음과 같이 업데이트합니다. $\theta_{t+1} = \theta_{E/} - \alpha \cdot x_{/E/}$ $_{i=1}$ $x_{/2t}$. 우리가 선택한다면 β_2 되려고

아래에서 무한히 1에 가깝다. 임 $_{\beta}$ $_{i=1}$ $x_{/2t}$. AdaGrad는
아담 버전 $\beta_1 = 0$, 무한 (1 $-_{2\to 1}$ $V_t = E/_{-1} \cdot \sum_{E/}$ $\beta_{2)}$ 및 대체 α 에 의해 $\sqrt{}$ n 어닐링 된 버전

$\alpha_t = \alpha \cdot E/_{-1/} \sqrt{}$ ^{2,} $\sum$ 즉 $\theta_{E/} - \alpha \cdot E/_{-1/2} \cdot$ 미/디/엄 $_{E//}$임 $\sqrt{}$ $_{\beta_2 \to 1}$ $V_t = \theta_{E/} - \alpha \cdot -/ \cdot$ $E/_{12}$ $g/t_{-1} \cdot \sum_{E/}$ $_{i=1}$ $x_{/2t}$ =

$\theta_{E/} - \alpha \cdot x_{/E//}$ $_{E/}$ $_{i=1}$ $x_{/2t}$. Adam과 Adagrad 간의 이러한 직접적인 통신은
편향 수정 항을 제거 할 때 유지되지 않습니다. RMSProp 에서처럼 바이어스 보정없이 β_2
무한히 1에 가까울수록 바이어스가 극도로 커지고 매개 변수 업데이트가 극히 커집니다.

6 E XPERIMENTS

제안 된 방법을 실증적으로 평가하기 위해 로지스틱 회귀, 다층 완전 연결 신경망 및 심층 컨벌루션 신경망을 포함한 다양한 인기 기계 학습 모델을 조사했습니다. 대규모 모델과 데이터 세트를 사용하여 Adam이 실용적인 딥 러닝 문제를 효율적으로 해결할 수 있음을 보여줍니다.

다른 최적화 알고리즘을 비교할 때 동일한 매개 변수 초기화를 사용합니다. 학습률 및 운동량과 같은 하이퍼 파라미터는 조밀 한 그리드에서 검색되고 결과는 최상의 하이퍼 파라미터 설정을 사용하여보고됩니다.

6.1 E XPERIMENT : 엘 OGISTIC 아르 자형 외출

MNIST 데이터 세트를 사용하여 L2 정규화 다중 클래스 로지스틱 회귀에 대해 제안 된 방법을 평가합니다. 로지스틱 회귀는 잘 연구 된 볼록 목표를 가지고있어 비교에 적합합니다.
걱정하지 않고 $\sqrt{}$ 지역 최소 문제에 대해. Stepize α 우리 물류에서
회귀 실험은 다음에 의해 조정됩니다. 1/ $E/$ 부패, 즉 $\bar{\alpha}_{t} = \sqrt{}$ $\frac{\alpha}{E/}$ 우리 이론과 일치하는
섹션 4의 ical 예측. 로지스틱 회귀는 784 차원 이미지 벡터에서 직접 클래스 레이블을 분류합니다. Adam을 Nesterov 모멘텀이있는 가속 SGD와 128의 미니 배치 크기를 사용하여 Adagrad와 비교합니다. 그림 1에 따르면 Adam이 모멘텀이있는 SGD와 유사한 수렴을 생성하고 둘 다 Adagrad보다 빠르게 수렴하는 것을 발견했습니다.

(Duchi et al., 2011)에서 논의 된 바와 같이, Adagrad는 희소 특성과 그라데이션을 효율적으로 처리 할 수 있습니다.
ent s는 주요 이론적 결과 중 하나이지만 SGD는 희귀 기능을 학습하는 데 부족합니다. 아담과
1/ $E/$ 이론적으로 Adagrad의 성능과 일치해야 stepize의 붕괴. (Maas et al., 2011)의 IMDB 영화 리뷰 데이터 세트를 사용하여 희소 특징 문제를 조사합니다. IMDB 영화 리뷰를 처음 10,000 개의 가장 자주 사용되는 단어를 포함하는 BoW (bag-of-words) 특징 벡터로 사전 처리합니다. 각 리뷰에 대한 10,000 차원 BoW 특징 벡터는 매우 희소합니다. (Wang & Manning, 2013)에서 제안한대로 50 % 드롭 아웃 노이즈가 BoW 기능에 적용될 수 있습니다.

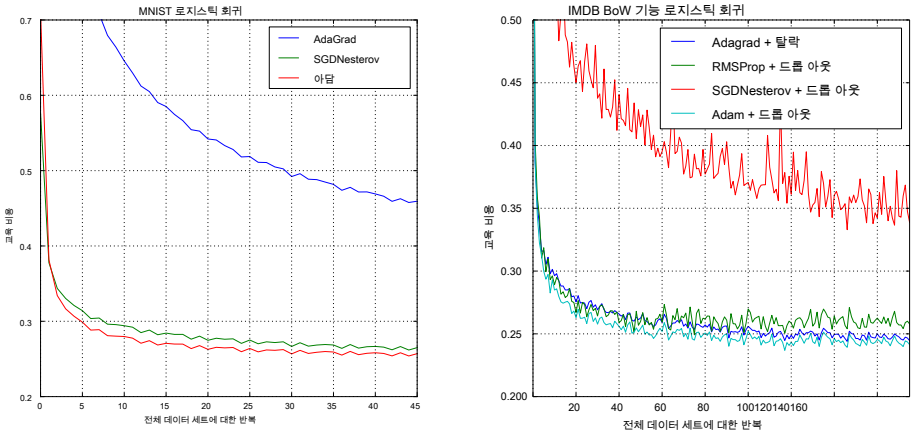


그림 1 : 10,000 개의 BoW (bag-of-words) 특징 벡터를 사용하여 MNIST 이미지 및 IMDB 영화 리뷰에 대한 음의 로그 가능성을 훈련시키는 로지스틱 회귀.

과적 합을 방지하기위한 훈련. 그림 1에서 Adagrad는 드롭 아웃 잡음 유무에 관계없이 Nesterov 모멘텀으로 SGD를 능가합니다. Adam은 Adagrad만큼 빠르게 수렴합니다. Adam의 경험적 성능은 섹션 2 및 4의 이론적 결과와 일치합니다. Adagrad와 유사하게 Adam은 희소 특성을 활용하고 운동량이있는 일반 SGD보다 빠른 수렴 속도를 얻을 수 있습니다.

6.2 E XPERIMENT : 미디엄 ULTI-LAYER 엔 EURAL 엔 ETWORKS

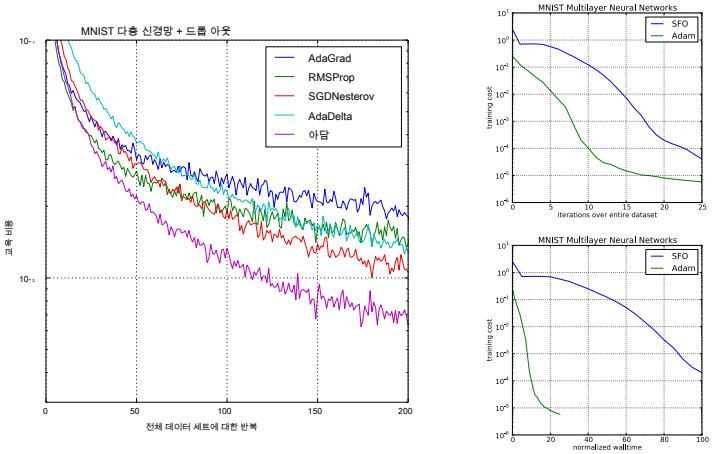
다층 신경망은 비 볼록 목적 함수가있는 강력한 모델입니다. 우리의 수렴 분석은 비 볼록 문제에는 적용되지 않지만 경험적으로 Adam은 이러한 경우 다른 방법보다 성능이 더 우수하다는 것을 경험적으로 발견했습니다. 실험에서 우리는 해당 지역의 이전 출판물과 일치하는 모델을 선택했습니다. 각각 1000 개의 은닉 유닛이있는 2 개의 완전히 연결된 은닉층이있는 신경망 모델과 ReLU 활성화가 미니 배치 크기가 128 인이 실험에 사용됩니다.

첫째, 표준 결정 론적 교차 엔트로피 목적 함수를 사용하여 다양한 최적화 프로그램을 연구합니다.
와 함께 ℓ_2 과적 합을 방지하기 위해 매개 변수의 무게 감소. SFO (sum-of-functions) 방법 (Sohl-Dickstein et al., 2014)은 최근 제안 된 준 뉴턴 방법으로
데이터를 최소화하고 다층 신경망의 최적화에서 좋은 성능을 보여주었습니다. 우리는 그들의 구현을 사용하고 그러한 모델을 훈련시키기 위해 Adam과 비교했습니다. 그림 2는 Adam이 반복 횟수와 벽시계 시간 측면에서 더 빠른 진행을 보여줍니다. 곡률 정보를 업데이트하는 데 드는 비용으로 인해 SFO는 Adam에 비해 반복 당 5-10 배 더 느리고 메모리 요구 사항은 미니 배치 수에서 선형 적입니다.

드롭 아웃과 같은 확률 적 정규화 방법은 오 버핏을 방지하는 효과적인 방법이며 단순성으로 인해 실제로 자주 사용됩니다. SFO는 결정 론적 하위 기능을 가정하고 실제로 확률 적 정규화로 비용 함수에 수렴하지 못했습니다. Adam의 효과를 드롭 아웃 노이즈로 훈련 된 다층 신경망에 대한 다른 확률 적 1 차 방법과 비교합니다. 그림 2는 우리의 결과를 보여줍니다. Adam은 다른 방법보다 더 나은 수렴을 보여줍니다.

6.3 E XPERIMENT : 씨 온볼 루션 엔 EURAL 엔 ETWORKS

컨볼 루션, 풀링 및 비선형 단위의 여러 계층이있는 컨볼 루션 신경망 (CNN)은 컴퓨터 비전 작업에서 상당한 성공을 거두었습니다. 대부분의 완전히 연결된 신경망과 달리 CNN의 가중치 공유는 서로 다른 레이어에서 매우 다른 그라디언트를 생성합니다. 합성 곱 계층에 대한 더 작은 학습률은 SGD를 적용 할 때 실제로 사용되는 경우가 많습니다. 우리는 깊은 CNN에서 Adam의 효과를 보여줍니다. 우리의 CNN 아키텍처는 5x5 컨볼 루션 필터와 3x3 최대 풀링의 세 번의 번갈아 가며 2 단계의 보폭을 가진 다음 1000 개의 ReLU (정밀 선형 은닉 유닛)의 완전히 연결된 계층이 뒤 따릅니다. 입력 이미지는 미백으로 전처리되며



(ㅏ) (비)

그림 2 : MNIST 이미지에 대한 다층 신경망 훈련. (a) 드롭 아웃 확률 적 정규화를 사용하는 신경망. (b) 결정 론적 비용 함수를 가진 신경망. SFO (sum-of-functions) 옵티 마이저 (Sohl-Dickstein et al., 2014)와 비교합니다.

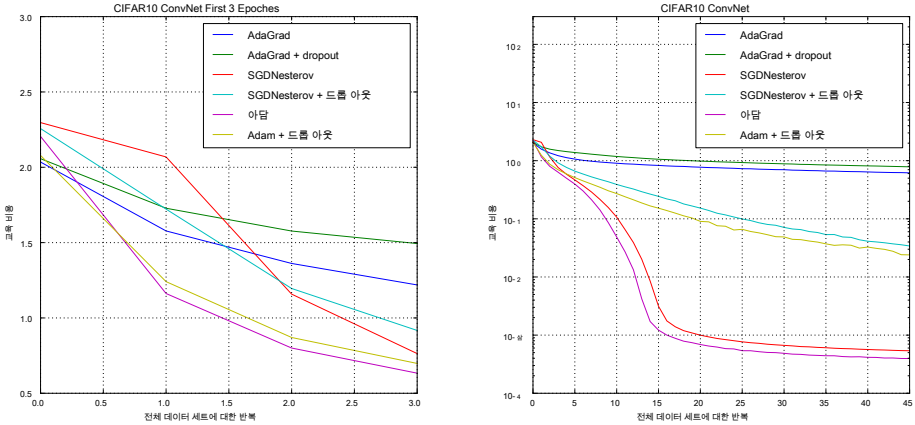
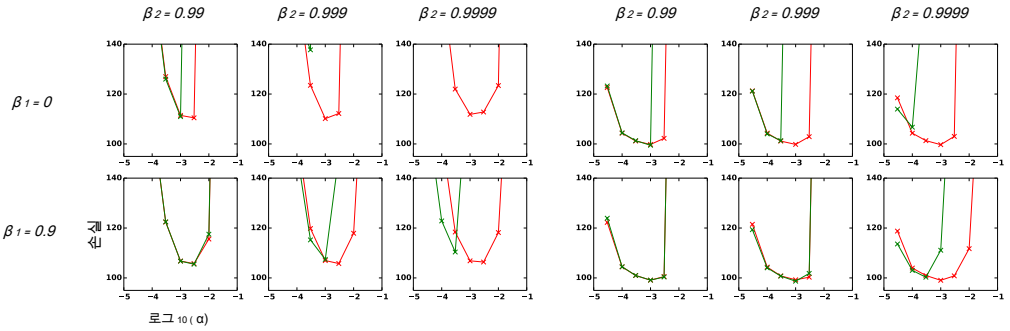


그림 3 : 컨볼 루션 신경망 훈련 비용. (왼쪽) 처음 세 시대의 교육 비용. (오른쪽) 45 epoch 이상의 교육 비용. c64-c64-c128-1000 아키텍처를 사용하는 CIFAR-10.

드롭 아웃 노이즈는 입력 레이어와 완전 연결 레이어에 적용됩니다. 미니 배치 크기도 이전 실험과 유사하게 128로 설정됩니다.

흥미롭게도 Adam과 Adagrad는 그림 3 (왼쪽)에 표시된 것처럼 교육 초기 단계에서 비용을 낮추면서 빠르게 진행되지만 Adam과 SGD는 결국 상당히 수렴합니다.
그림 3 (오른쪽)에 표시된 CNN의 경우 Adagrad보다 빠릅니다. 두 번째 모멘트 추정 V_{t}
몇 epoch 후에 0으로 사라지고 ε 알고리즘 1에서 두 번째 순간
따라서 추정치는 6.2 절의 완전히 연결된 네트워크와 비교하여 CNN의 비용 함수 기하학에 대한 잘못된 근사치입니다.
반면, 첫 번째 순간까지 미니 배치 분산을 줄이는 것이 CNN에서 더 중요하며 속도 향상에 기여합니다. 결과적으로 Adagrad는이 특정 실험에서 다른 것보다 훨씬 느리게 수렴합니다. Adam은 모멘텀으로 SGD보다 약간의 개선을 보이지만 SGD에서와 같이 수동으로 선택하는 대신 여러 계층에 대한 학습률 척도를 조정합니다.



(a) 10 세대 후 (b) 100 epoch 이후

그림 4 : Variational Auto-Encoder를 학습 할 때 손실 (y 측)에 대한 10 epoch (왼쪽) 및 100 epoch (오른쪽) 후 바이어스 보정 항 (빨간색 선) 대 바이어스 보정 항 없음 (녹색 선)의 효과 VAE) (Kingma & Welling, 2013), stepize의 다양한 설정 α (x 측) 및 하이퍼
매개 변수 β_1 과 β_2.

6.4 E XPERIMENT : 편향 수정 기간

또한 섹션 2와 3에 설명 된 편향 보정 항의 효과를 경험적으로 평가합니다. 섹션 5에서 논의 된 편향 보정 항을 제거하면 RMSProp (Tiele-
man & Hinton, 2012). 우리는 다양합니다 β_1 과 β_2 단일 솜김을 사용하여 (Kingma & Welling, 2013)과 동일한 아키텍처로 VAE (변형 자동 인코더)를 훈련 할 때
소프트 플러스 비선형 성과 50 차원 구형 가우스 잠재 변수를 가진 500 개의 은닉 유닛이있는 레이어. 광범위한 하이퍼 매개 변수 선택을 반복했습니다. $\beta_1 \in [0, 0.9]$ 과
$\beta_2 \in [0.99, 0.999, 0.9999]$, 과 로그 $_{10}$ (α) $\in [-5, ..., -1]$. 가치 β_2 1에 가까우며, 희소 기울기에 대한 견고성에 필요하므로 초기화 편향이 더 커집니다. 따라서 우리는 편향 보정을 기대합니다
이러한 느린 붕괴의 경우 용어는 최적화에 대한 악영향을 방지하는 데 중요합니다.

그림 4에서 값 β_2 1에 가까워지면 편향 수정 항이 없을 때, 특히 훈련의 처음 몇 세대에서 훈련이 불안정 해집니다. 최고의 결과는 작은
값 (1 − β_2) 및 편향 보정; 이것은 은닉 유닛이 특정 패턴에 특화되어 기울기가 희박 해지는 경향이있을 때 최적화가 끝날 무렵 더욱 분명해졌습니다. 요약하면, Adam
하이퍼 파라미터 설정에 관계없이 RMSProp와 같거나 더 나은 성능을 제공합니다.

7 E XTENSIONS

7.1A DA 미디엄 도끼

Adam에서 개별 가중치에 대한 업데이트 규칙은 (스케일 된) $엘_2$ 개별 현재 및 과거 그라디언트의 규범. 우리는 일반화 할 수 있습니다 $엘_2$ 표준 기반 업데이트 규칙을 $엘_피$ 표준 기반 업데이트 규칙. 이러한 변종은 대규모의 경우 수치 적으로 불안정 해집니다.
$피$. 그러나 특별한 경우에는 $피 \to \infty$, 놀랍도록 간단하고 안정적인 알고리즘이 등장합니다. 알고리즘 2를 참조하십시오.
이제 알고리즘을 유도합니다. 의 경우하자 $엘_피$ 규범, 단계적
시간에 $티$ 반비례하다 $V_{1/피}$ $_t$, 어디:

$$V_t = \beta_2^{피} V_{티-1} + (1 - \beta_2^{피}) |x_t|_p^{피} \tag{6}$$

$$= (1 - \beta_2^{피}) \sum_{i=1}^{티} \beta_2^{(t-나는)} \cdot |x_i|_p^{피} \tag{7}$$

알고리즘 2 : *AdaMax,* 무한도 규범에 기반한 아담의 변형. 자세한 내용은 섹션 7.1을 참조하십시오.
테스트 된 기계 학습 문제에 대한 좋은 기본 설정은 다음과 같습니다. $\alpha = 0.002$, $\beta_1 = 0.9$ 과
$\beta_2 = 0.999$. 와 β_t 우리는 β_1 권력에 t. 여기, $\alpha / (1 - \beta_t)$ 편향 수정 항이있는 학습률입니다. 1 처음으로. 벡터에 대한 모든
연산 a 1 요소별로 다시.

필요 : α : 스텝 사이즈
필요 : $\beta_1, \beta_2 \in [0, 1)$: 지수 붕괴율
필요 : $f(\theta)$: 매개 변수가있는 확률 적 목적 함수 θ
필요 :
$m_0 \leftarrow 0$ (1st 초기화 매개 변수 벡터)
$u_0 \leftarrow 0$ (지수 가중치가 적용된 무한대 규범 초기화)
t 0 (타임 스텝 초기화)
동안 θ_t 수렴되지 않음 하다
$t \leftarrow t + 1$
$g_t \leftarrow \nabla$
$m_t \leftarrow \beta$ (업데이트 1st 시간 편향 모멘트 추정)
$u_t \leftarrow$ 최대 ($\beta_2 \cdot u_{t-1}$, $|g_t|$) (지수 가중치 무한대 기준 업데이트)
θ_t θ_{t-1} ($\alpha / (1 - \beta_1^t)) \cdot m_t / u_t$ 매개 변수 업데이트)
동안 끝
반환 θ_t (결과 매개 변수)

여기서 붕괴 항은 다음과 같이 동등하게 매개 변수화됩니다. β_p 2 대신에 β_2. 이제 $p \to \infty$,
그리고 정의 u_t 임 $p \to \infty$ (V_p) 1/ p 그때:

$$u_t = \lim_{p \to \infty} (V_p)^{1/p} = \lim_{p \to \infty} \left(\sum_{i=1}^{t} \beta_2^{p(t-i)} \cdot |g_i|^p\right)^{1/p} \tag{8}$$

$$= \lim_{p \to \infty} (1 - \beta_2^2) \left(\sum_{i=1}^{t} \beta_2^{p(t-i)} \cdot |g_i|^p\right)^{1/p} \tag{9}$$

$$= \lim_{p \to \infty} \left(\sum_{i=1}^{t} \beta_2^{p(t-i)} \cdot |g_i|^p\right)^{1/p} \tag{10}$$

$$= \max(\beta_2^{t-1} |g_1|, \beta_2^{t-2} |g_2|, \ldots, \beta_2 |g_{t-1}|, |g_t|) \tag{11}$$

매우 간단한 재귀 공식에 해당합니다.

$$u_t = \max(\beta_2 \cdot u_{t-1}, |g_t|) \tag{12}$$

초기 값으로 $u_0 = 0$. 편리하게도 초기화를 위해 수정할 필요가 없습니다.
이 경우 편향. 또한 t | hat | th ≤ 매개 변수 업데이트의 크기는 Adam보다 AdaMax와 더 간단합니다. Δ_t
α.

7.2 톤 EMPORAL AVERAGING

마지막 반복 이후 ∑ 확률 적 근사로 인해 잡음이있는 경우 일반화 성능이 향상됩니다.
종종 평균화에 의해 달성됩니다. 이전 Moulines & Bach (2011)에서 Polyak-Ruppert 평균화 (Polyak & Juditsky, 1992;
Ruppert, 1988)는 표준 수렴을 개선하는 것으로 나타났습니다.
SGD, 여기서 θ_{t+1} $\bar{\theta}_t = \frac{1}{t} \sum_{k=1}^{t} \theta_k$ 또는 매개 변수에 대한 지수 이동 평균은
최신 매개 변수 값에 더 높은 가중치를 부여하여 사용
알고리즘 1과 2의 내부 루프에 한 줄을 추가합니다. $\bar{\theta}_t \leftarrow$ 에스. 일 · 사소하게 구현 될 수 있음 (있음 β_2 2). 변형은 추정기에 의해
다시 수정 될 수 있습니다. $\theta = \theta / (1 - \beta_t)$

8C 온 클루 션

확률 적 목적 함수의 기울기 기반 최적화를위한 간단하고 계산적으로 효율적인 알고리즘을 도입했습니다. 우리의 방법은
다음과 같은 기계 학습 문제를 목표로합니다.

대규모 데이터 세트 및 / 또는 고차원 매개 변수 공간. 이 방법은 최근에 널리 사용되는 두 가지 최적화 방법의 장점, 즉 희소 기울기를 처리하는 AdaGrad 기능과 고정되지 않은 목표를 처리하는 RMSProp 기능을 결합합니다. 이 방법은 구현이 간단하고 메모리가 거의 필요하지 않습니다. 이 실험은 볼록 문제에서 수렴 률에 대한 분석을 확인합니다. 전반적으로 Adam은 필드 머신 러닝의 광범위한 비 볼록 최적화 문제에 견고하고 적합하다는 것을 알았습니다.

9A CKNOWLEDGMENTS

이 문서는 Google Deepmind의 지원 없이는 존재하지 않았을 것입니다. Adam이라는 이름을 만든 Ivo Danihelka와 Tom Schaul에게 특별한 감사를드립니다. 원래 AdaMax 파생에서 오류를 발견 한 Duke University의 Kai Fan에게 감사드립니다. 이 작업의 실험은 부분적으로 SURF 재단의 지원을 받아 네덜란드 국가 전자 인프라에서 수행되었습니다. Diederik Kingma는 Google European Doctorate Fellowship in Deep Learning의 지원을받습니다.

아르 자형 EFERENCES

아마리, 슌이치. 자연 그라디언트는 학습에서 효율적으로 작동합니다. *신경 계산,* 10 (2) : 251–276, 1998.

Deng, Li, Li, Jinyu, Huang, Jui-Ting, Yao, Kaisheng, Yu, Dong, Seide, Frank, Seltzer, Michael, Zweig, Geoff, 그는 Xiaodong, Williams, Jason, et al. Microsoft의 음성 연구를위한 딥 러닝의 최근 발전. *ICASSP 2013,* 2013.

Duchi, John, Hazan, Elad 및 Singer, Yoram. 온라인 학습 및 확률론을위한 적응 형 하위 구배 방법 최적화. *기계 학습 연구 저널,* 12 : 2121–2159, 2011.

그레이브스, 알렉스. 순환 신경망을 사용하여 시퀀스 생성. *arXiv 프리 프린트 arXiv : 1308.0850,* 2013. Graves, Alex, Mohamed,

Abdel-rahman, Hinton, Geoffrey. 심층 반복 신경을 사용한 음성 인식 네트워크. 에 *음향, 음성 및 신호 처리 (ICASSP), 2013 IEEE International Conference on,* 6645–6649 쪽. IEEE, 2013.

Hinton, GE 및 Salakhutdinov, RR 신경망으로 데이터의 차원을 줄입니다. *과학,* 313 (5786) : 504–507, 2006.

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N, et al. 음성 인식의 음향 모델링을위한 심층 신경망 : 4 개 연구 그룹의 공유 된 견해. *Signal Processing Magazine, IEEE,* 29 (6) : 82–97, 2012a.

Hinton, Geoffrey E, Srivastava, Nitish, Krizhevsky, Alex, Sutskever, Ilya 및 Salakhutdinov, Ruslan R. Im- 기능 탐지기의 공동 적응을 방지하여 신경망을 증명합니다. *arXiv 프리 프린트 arXiv : 1207.0580,* 2012b.

Kingma, Diederik P 및 Welling, Max. 자동 인코딩 Variational Bayes. 에 *제 2 회 국제 학습 대회 학습 표현 (ICLR)에 대해,* 2013.

Krizhevsky, Alex, Sutskever, Ilya 및 Hinton, Geoffrey E. Imagenet 분류 신경망. 에 *신경 정보 처리 시스템의 발전,* 1097–1105, 2012 쪽.

Maas, Andrew L, Daly, Raymond E, Pham, Peter T, Huang, Dan, Ng, Andrew Y 및 Potts, Christopher. 감정 분석을위한 단어 벡터 학습. 에 *전산 언어학 협회 제 49 차 연례 회의 : 인간 언어 기술-제 1 권,* 142–150 쪽. 협회 전산 언어학, 2011.

Moulines, Eric and Bach, Francis R. 다음에 대한 확률 적 근사 알고리즘의 비 점근 분석 기계 학습. 에 *신경 정보 처리 시스템의 발전,* 451–459, 2011 쪽.

Pascanu, Razvan 및 Bengio, Yoshua. 딥 네트워크를위한 자연스러운 그래디언트 재 방문. *arXiv 프리 프린트 arXiv : 1301.3584,* 2013.

Polyak, Boris T 및 Juditsky, Anatoli B. 평균화에 의한 확률 적 근사 가속. *SIAM 저널 제어 및 최적화에 대해* 30 (4) : 838–855, 1992 년.

Roux, Nicolas L 및 Fitzgibbon, Andrew W. 빠른 자연 뉴턴 방법. 에 *27 일 회의록 기계 학습에 관한 국제 회의 (ICML-10),* 623–630, 2010 쪽.

루퍼트, 데이비드. 천천히 수렴하는 robbins-monro 프로세스로부터의 효율적인 추정. 기술 보고서, 코넬 대학 운영 연구 및 산업 공학, 1988.

Schaul, Tom, Zhang, Sixin 및 LeCun, Yann. 더 이상 성가신 학습률이 없습니다. *arXiv 프리 프린트 arXiv : 1206.1106,* 2012.

Sohl-Dickstein, Jascha, Poole, Ben 및 Ganguli, Surya. stochas를 통합하여 빠른 대규모 최적화- tic gradient 및 quasi-newton 방법. 에 *제 31 회 기계 학습에 관한 국제 컨퍼런스 (ICML-14) 절차,* 604–612, 2014 쪽.

Sutskever, Ilya, Martens, James, Dahl, George 및 Hinton, Geoffrey. 초기화의 중요성과 딥 러닝의 추진력. 에 *기계 학습에 관한 제 30 회 국제 컨퍼런스의 진행 (ICML-13),* 1139–1147, 2013 쪽.

Tieleman, T. and Hinton, G. 강의 6.5-RMSProp, COURSERA : 기계 학습을위한 신경망. 기술 보고서, 2012.

왕,시다, 매닝, 크리스토퍼. 빠른 중퇴 교육. 에 *제 30 회 국제 학술 대회 절차- 따라서 기계 학습 (ICML-13)에서* 118–126, 2013 쪽.

Zeiler, Matthew D. Adadelta : 적응 형 학습률 방법. *arXiv 프리 프린트 arXiv : 1212.5701,* 2012. Zinkevich, Martin. 온라인 볼록

프로그래밍 및 일반화 된 무한 경사 상승. 2003.

10A 부록

10.1 C ONVERGENCE 피 지붕

정의 10.1. *기능 f : R 디/ → 아르 자형 모두를 위해 볼록하다 x, y ∈ 아르 자형 디/, 모든 λ ∈ [*0, 1],

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(\lambda x + (1 - \lambda) y)$$

또한 볼록 함수는 접선에서 초평면에 의해 하한이 될 수 있습니다.

정리 10.2. *기능 f : R 디/ → 아르 자형 볼록한 다음 모두 x, y ∈ 아르 자형 디/,*

$$f(y) \geq \text{에프 엑스} + \nabla \text{에프 엑스}_{T(} \text{와이} - \text{엑스})$$

위의 기본형은 후회를 상한선으로 사용하는 데 사용할 수 있으며 주 정리에 대한 우리의 증명은 초평면을 Adam 업데이트 규칙으로 대체하여 구성됩니다.

다음 두 가지 기본 정리는 우리의 주요 정리를 지원하는 데 사용됩니다. 또한 몇 가지 정의 시뮬레이션을 사용합니다.
우리의 표기법을 만족 시키십시오. $지/_{t, \nabla \text{에프 } 티(\theta 티)}$과 $지/_{t, i}$로 *나는*일 요소. 우리는 정의 $지/_{1: t, i \in}$ 아르 자형 티/포함하는 벡터로 *나는*일 모든 반복에 대한 그래디언트의 차원 *t*, $g_{1: t, i = [} 지/_{1, \text{나는}}, 지/_{2, \text{나는}}, \cdots, 지/_{t, i]}$

정리 10.3. *허락하다* $지/_{t = \nabla \text{에프 } 티(\theta 티)}$ *an* $\sum\sqrt{}$ *디* $지/_{1: 티/}$ *위와 같이 정의되고 제한되어야합니다.* $\| 지/_{티/} \|_2 \leq 지, \| 지/_{티/} \|_\infty \leq 지/_\infty$. *그때,*

$$\sum_{t=1}^{티/} \sqrt{\frac{지/_{2}_{t, i}}{티/}} \leq 2 지/_\infty \| 지/_{1: T, i} \|_2$$

증명. 우리는 불평등을 증명할 것입니다 √ y T에 대한 유도를 사용합니다.

기본 케이스 $T = 1$, 우리는 $지/_2$ $\sqrt{}$ $_{1, \text{나는}} \leq 2 지/_\infty \| 지/_{1, \text{나는}} \|_2$.

귀납적 단계의 경우

$$\sum_{t=1}^{\text{팀}} \sqrt{\frac{지/_{2}_{t, i} =}{티/}} \sum_{t=1}^{티/} \sqrt{\frac{지/_{2}_{t, i +}}{티/}} \sqrt{\frac{지/_{2}_{T, i}}{티/}}$$

$$\leq 2 지/_\infty \| \sqrt{} 지/_{1: 티/ - 1, \text{나는}} \|_2 + \sqrt{\frac{지/_{2}_{T, i}}{티/}}$$

$$= 2 지/_\infty \| 지/_{1: T, i} \|_2 \sqrt{\| 지/_{1: T, i} \|_2^2 - 지/_{티/}^2} + \sqrt{\frac{지/_{2}_{T, i}}{티/}}$$

에서, $\| 지/_{1: T, i} \|_2 - 지/_{2\ T, i} + \sqrt{\frac{지/_{t, i}}{4 \| 지/_{1: T, i} \|_2}} \geq \| 지/_{1: T, i} \|_2 - 지/_{2\ T, i}$, 우리는 양쪽의 제곱근을 취할 수 있고 있다,

$$\sqrt{\| 지/_{1: T, i} \|_2^2 - 지/_{2\ T, i}} \leq \| 지/_{1: T, i} \|_2 - \frac{지/_{2}_{T, i}}{2 \| \sqrt{} 지/_{1: T, i} \|_2}$$

$$\leq \| 지/_{1: T, i} \|_2 - \frac{지/_{2}_{T, i}}{2 \sqrt{TG_2}_\infty}$$

불평등을 재정렬하고 √ 대체 $\| 지/_{1: \sqrt{}} \|_2$ $\sqrt{\| 지/_{T, i} \|_2 - 지/_{2\ T, i}}$기간,

$$지/_\infty \sqrt{\| 지/_{1: T, i} \|_2^2 - 지/_{2\ T +}} \sqrt{\frac{지/_{2}_{T, i}}{티/}} \leq 2 지/_\infty \| 지/_{1: T, i} \|_2$$

□

정리 10.4. *허락하다 ,* $\sqrt{\gamma}$ $\frac{1}{\beta}$ *에 대한* $\beta_1, \beta_2 \in [0, 1)$ *만족하는* $\sqrt{\beta_1} < 1$ *그리고* $\frac{2}{\beta_2}$ *경계 지* $_{티,}$ $\|$ *지* $_{티}\| \leq$ *지,* 2 $\|$ *지* $_{티}\|_\infty \leq$ *지* $_\infty$*, 후속* $_2$ *미균 병사* $\sum$ *미디엄* $_{t,i}$ *부정적 보유*

$$\sum_{t=1}^{티} \frac{\sqrt{2}}{TV_{t,i}} \quad \frac{}{1\ \gamma\ 1\sqrt{1-\beta_2}} \quad \|지_{1:T,i}\|_2$$

증명. 가정하에 $\frac{\sqrt{1\ \beta_{티}} \leq}{(1\ \beta)\ _{티/2}\sqrt{(1-\beta)_2}.\sum} \frac{1}{1}$ 요약에서 마지막 용어를 확장 할 수 있습니다.
사용 $\sum$ upd $\sqrt{}$..에서 ê 알고리즘의 규칙 $_1$ 1,

$$\sum_{t=1}^{티} \frac{미디엄^2_{t,T}}{TV_{t,i}} = \sum_{t=1}^{티-1} \frac{\sqrt{2}\ 미디엄_{t,T}}{\sqrt{TV_{t,i}}} + \frac{\sqrt{1-\beta_{티}^2}}{(1\ \beta_{T})^2} \frac{(\sum_{케이}(1-\beta_1)\ \beta_1^{티-케이}\ 지_{k,i})^2}{\sqrt{\sum_{j=1}(1-\beta_2)\ \beta_2^{티-제이}\ 지^2_{j,i}}}$$

$$\leq \sum_{t=1}^{티-1} \frac{미디엄^2_{t,i+}}{} \frac{\sqrt{1-\beta_{티}^2}}{(1\sqrt{}-\beta_{T})^2_1} \sum_{k=1}^{티} 티 \frac{\sqrt{티((1-\beta)\ \beta^{티-케이}_1\ 지_{k,i})^2}}{\sqrt{티 \sum_{j=1}^{T}(1-\beta_2)\ \beta_2^{티-제이}\ 지^2_{j,i}}}$$

$$\leq \sum_{t=1}^{티-1} \frac{미디엄^2_{t,i+}}{\sqrt{TV_{t,i}}} \frac{\sqrt{1-\beta_{티}^2}}{(1-\beta_{T})^2_1} \sum_{k=1}^{티} \frac{\sqrt{티((1-\beta)\ \beta^{티-케이}_1\ g_{k,i})^2}}{티(1-\beta_2)\ \beta_2^{티(-케이}\ 지^2\ 전부-케이\ \|지_{k,i}\|_2}$$

$$\leq \sum_{t=1}^{티-1} \frac{미디엄^2_{t,i+}}{\sqrt{티}\ TV} \frac{\sqrt{1\ \beta^2_{2티}}}{(1-\beta_{T})^2_1} \sqrt{\frac{(1\ \beta_1)^2}{티(1-\beta_2)}} \sum_{k=1}^{티} 티 \left(\frac{\sqrt{\beta_1}}{\beta_2}\right)^2$$

$$\leq \sum_{t=1}^{티-1} \frac{미디엄^2_{t,i}\ TV_{t,i}}{TV_{t,i}} \frac{}{티(1\ \beta_2)} \sum_{k=1}^{티} \gamma^{티-케이}\ \|지_{k,i}\|_2$$

마찬가지로, 우리는 $\sum$ 그만큼의 rêst $\sum$ 합계의 용어.

$$\sum_{t=1}^{티} \frac{미디엄^2_{t,T}}{TV_{t,i}} \leq \sum_{t=1}^{티} \frac{\|지_{t,i}\|_2}{\sqrt{티(1-\beta_2)}} \sum_{j=0}^{티-티} t\gamma^{제이}$$

$$\leq \sum_{t=1}^{티} \frac{\|지_{t,i}\|_2}{티(1-\beta_2)} \sum_{j=0}^{티} t\gamma^{제이}$$

에 대한 $\gamma < 1$, 사용 $\sum\sqrt{}\ \|지_{t,i}\|_2$의 상한 $\sum$ 그는 산술 기하학 시리즈 $\sum t\gamma^{t} < \frac{1}{(1\ \gamma)^2}$에스, $\sum_{티}$ $\frac{}{-1}$

$$\sum_{t=1}^{티} \frac{}{티(1-\beta_2)} \sum_{j=0}^{티} t\gamma^{제이} \leq \frac{1}{(1-\gamma)^2\sqrt{1-\beta_2}} \sum_{t=1}^{티} \frac{\|지_{t,i}\|_2}{\sqrt{티}}$$

Lemma 10.3 적용,

$$\sum_{t=1}^{티} \frac{미디엄^2_{t,i}}{\sqrt{TV_{t,i}}} \leq \frac{2\ 지_\infty}{(1-\gamma)^2\sqrt{1-\beta_2}} \|지_{1:T,i}\|_2$$

□

표기법을 단순화하기 위해 γ , $\frac{\sqrt{\beta_1^2}}{\beta_2}$ 직관적으로, 우리의 다음 정리는

학습률 $\alpha_{티}$ 속도로 쇠퇴하고 있습니다 $티_{-1}$ $_2$ 평균 계수를 실행하는 첫 번째 순간 $\beta_{1,티}$ 부식

기하 급수적으로 λ, 일반적으로 1에 가깝습니다. 예 : $1-10^{-8}$.

정리 10.5.

지 $_\infty$ *모든* $\theta \in$ *함수가 에프 티 경계가 있고 그라디언트가 있습니다 Adam에 의해 생성 된, 에프 (θ)* $\|_2 \leq$ *디,*

" θ 미디엄 − θ 엔 " ∞ ≤ 디/ ∞ 어떠한 것도 m, n ∈ { 1, ..., 티}, 과 β 1, β 2 ∈ [0, 1) 폴다 √ β 2 / 1 < 1. 허락하다 α t = √ α / 티/ 과 β 1, t = β 1 λ 티/ − 1, 람다 ∑ 는 (0 √ 1). Adam은 다음 구를 달성합니다. ∑ 모두를 위해 ∑ 디/ ∞ 티/ ≥ 1.

$$R(T) \le \frac{D_\infty^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{D} \sqrt{T\hat{v}_{T,i}} + \frac{\alpha(\beta_1+1)\sqrt{1-\beta_2}}{(1-\beta_1)(1-\gamma)^2} \sum_{i=1}^{D} \|s_{1:T,i}\|_2 + \sum_{i=1}^{D} \frac{D_\infty^2 \sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2} \|s_{1:T,i}\|_2$$

증명. Lemma 10.2를 사용하여,

$$\text{에프 티/}(\theta_{\text{티/}}) - \text{에프 티/}(\theta^*) \le s_{\text{티/}}{}^t(\theta_{\text{티/}} - \theta^*) = \sum_{i=1}^{D} s_{t,i}(\theta_{t,i} - \theta^*_{,\text{나는}})$$

알고리즘에 제시된 업데이트 규칙에서 √ 1,

$$\theta_{t+1} = \theta_{\text{티/}} - \alpha_{\text{티/}} \text{미디엄}_t / (\sqrt{V_{\text{티/}}}) = \theta_{\text{티/}} - \alpha \frac{\beta \hat{}, \text{티/}\ \text{미디엄}_{-+} / (1-\beta_{\text{티/}})}{\sqrt{V_{\text{티/}}}}$$

우리는 *나는*일 매개 변수 벡터의 차원 $\theta_{\text{티/}} \in$ *아르 자형* 디 스칼라 빼기 θ^*위의 업데이트 규칙의 양쪽을 정사각형으로 만듭니다.

$$(\theta_{t+1,\text{나는}} - \theta^*_{,\text{나는}})^2 = (\theta_{t,i} - \theta^*_{,\text{나는}})^2 - 2(\theta_{t,i} - \theta^*_{,\text{나는}}) \frac{\alpha \hat{} \text{미디엄}_{t,i}}{\sqrt{V_{t,i}}} + \frac{\beta_{1,\text{티/}}}{(1-\beta_{1,\text{티/}})\sqrt{V_{t,i}}} (\theta_{t,i} - \theta^*_{,\text{나는}})^2 + \alpha_t^2 \frac{\text{미디엄}_{t,i}^2}{V_{t,i}}$$

우리는 재정렬 할 수 있습니다 √ abOve 방정식과 사용 Yo √ ung의 불평등, $ab \le a^2/2 + b^2/2$. 또한 그것을 보여 주었다 $\hat{v}_{t,i} = \sum_{j=1}^{\text{티/}} (1-\beta_{2)}\beta_{\text{티/}}{}^{\text{티/}-j} s_{j,i}^2 / \sum_{\text{티/}} 1-\beta_{\text{티/}} \le \|s_{1):t,i}\|_2^2$ 과 $\beta_{1,\text{티/}} \le \beta_1$. 그때

$$g_{t,i} = \theta_{t,i} - \theta^*_{,\text{나는}} \quad \frac{(1-\beta_{\text{티/}1})\sqrt{V_{t,i}}}{2\alpha_{\text{티/}}(1-\beta_{\lambda,\text{티/}})} (\theta_{t,i} - \theta^*_{,\text{나는}})^2 - (\theta_{t+1,\text{나는}} - \theta^*_{,\text{나는}})^2$$

$$+ \frac{\beta_{1,\text{티/}}}{(1-\beta_{1,\text{티/}})\alpha_{\text{티/}-1}} \sqrt{V_{t-1,\text{나는}}} (\theta^* - \theta_{t,i}) \alpha \hat{}_{\text{티/}} \text{미디엄}_{-1,\text{나는}} \sqrt{V_{\text{티/}-1,\text{나는}}} + \frac{\alpha_{\text{티/}}(1-\beta_{1)} V_{t,i}{}^{\text{티/}} \text{미디엄}_{t,i}^2}{2(1-\beta_{1,\text{티/}})}$$

$$\le \frac{1}{2\alpha_{\text{티/}}(1-\beta_1)} (\sqrt{V_{t,i}}(\theta_{t,i} - \theta^*_{,\text{나는}})^2 - \sqrt{V_{\text{티/}1,\text{나는}}} (\theta_{t+1,\text{나는}} - \theta^*_{,\text{나는}})^2) + \frac{\beta_{1,\text{티/}}}{2\alpha_{\text{티/}}(1-\beta_{1,\text{티/}})} (\theta^* - \theta_{t,i})^2 \sqrt{V_{t,i}}$$

$$+ \frac{\beta_1 \alpha_{\text{티/}-1}}{2(1-\beta_1)(1-\beta_1)} \sqrt{V_{\text{티/}-1,\text{나는}}} \text{미디엄}_{\text{티/}-1,\text{나는}} + \frac{\alpha_{\text{티/}}}{2(1-\beta_1)} \sqrt{V_{t,i}} \text{미디엄}_{t,i}^2$$

Lemma 10을 적용합니다. 제위의 불평등에 대해 모두 합산 $\theta_{\text{티/}}) - f(\theta^*)$ 볼록 함수의 순서 ∑ ...에 대한 $\text{티/} 1, ..., T$:

$$R(T) \le D_\infty^2 \sum_{i=1}^{D} \frac{1}{2\alpha(1-\beta_1)} (\theta^*_{,\text{나는}} - \theta_{1,\text{나는}})^2 \sqrt{V_{1,i}} + \sum_{i=1}^{D} \sum_{t=2}^{T} \frac{1}{2(1-\beta_1)} (\theta_{t,i} - \theta^*_{,i})^2 \left(\frac{\sqrt{V_{t,i}}}{\alpha_t} - \frac{\sqrt{V_{\text{티/}1,\text{나는})}}}{\alpha_{\text{티/}-1}} \right)$$

$$+ \frac{\beta_1 \alpha G_\infty \sqrt{1-\beta_2}}{(1-\beta_1)(1-\beta_1)\sqrt{(1-\gamma)^2}} \sum_{i=1}^{D} \|s_{1:T,i}\|_2 \sqrt{T} + \frac{\alpha G_\infty}{(1-\beta_1)\sqrt{(1-\beta_2)(1-\gamma)^2}} \sum_{i=1}^{D} \|s_{1:T,i}\|_2$$

$$+ \sum_{i=1}^{D} \sum_{t=1}^{T} \sqrt{1^{\text{티/}}} - 2 \frac{\beta_1}{2\alpha_{\text{티/}}(1-\beta_{1),\text{나는}}} (\theta^*_{,\text{나는}} - \theta_{t,i})^2 \sqrt{V_{t,i}}$$

가정에서 ∑ , $\|\theta_{t-1} - \theta^*\|_2 \le C_1$, $\|\theta_{t,1} - \theta_t\|_\infty \le C_2\alpha_t$, 우리는 :

$$R(T) \le \frac{C_1}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sqrt{\hat{v}_{T,i}} + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2}\sum_{i=1}^{d}\|g_{1:T,i}\|_2 + \frac{C_2^2\sum_{t=1}^{T}\sum_{i=1}^{d}}{2\alpha}\frac{\beta_{1,t}}{(1-\beta_{1,t})}\sqrt{v_{t,i}}$$

$$\le \frac{C_1}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\frac{\hat{v}_{T,i}}{v_{T,i}} + \frac{\alpha(1+\beta_1)}{\sqrt{1-\beta_2}(1-\gamma)^2}\sum_{i=1}^{d}\|g_{1:T,i}\|_2$$

$$+ \frac{C_2^2\sqrt{1-\beta_2}}{2\alpha}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\beta_{1,t}}{(1-\beta_{1,t})}$$

산술 기하학을 사용할 수 있습니다. ∑ tric 시리즈 어퍼 보우 ∑ 마지막 학기 :

$$\sum_{t=1}^{T}\frac{\beta_{1,t}}{(1-\beta_{1,t})}\sqrt{t} \le \sum_{t=1}^{T}\frac{1}{(1-\beta_1)}\lambda^{t-1}\sqrt{t}$$

$$\le \sum_{t=1}^{T}\frac{1}{(1-\beta_1)}\lambda^{t-1}t$$

$$\le \frac{1}{(1-\beta_1)(1-\lambda)^2}$$

따라서 우리는 ∑ e fo 휴회하는 일 :

$$R(T) \le \frac{C_1}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sqrt{v_{T,i}} + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2}\sum_{i=1}^{d}\|g_{1:T,i}\|_2 + \sum_{i=1}^{d}\frac{C_2\|g_{1:T,i}\|_\infty\sqrt{1-\beta_2}}{2\alpha\beta_1(1-\lambda)^2}$$

□