ICLR 2015에서 컨퍼런스 논문으로 발표

아담: 확률적 최적화를 위한 방법

디에릭 P. 킹마*
암스테르담 대학교, OpenAI
dpkingma@openai.com

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

초록

저차 모멘트의 적응적 추정을 기반으로 확률적 목적 함수의 1차 기울기 기반 최적화를 위한 알고리즘인 *Adam을* 소개합니다. 이 방법은 구현이 간단하고 계산 효율이 높으며 메모리 요구량이 적고 기울기의 대각선 크기 조정에 변하지 않으며 데이터 및/또는 매개변수가 큰 문제에 적합합니다. 이 방법은 또한 고정되지 않은 목표와 매우 잡음이 많거나 희박한 그라데이션이 있는 문제에도 적합합니다. 하이퍼 파라미터는 직관적으로 해석할 수 있으며 일반적으로 튜닝이 거의 필요하지 않습니다. *아담이* 영감을 얻은 관련 알고리즘과의 연관성에 대해 설명합니다. 또한 알고리즘의 이론적 수렴 속성을 분석하고 온라인 볼록 최적화 프레임워크에서 가장 잘 알려진 결과와 유사한 수렴률에 대한 후회 한계를 제공합니다. 경험적 결과는 아담이 실제로 잘 작동하며 다른 확률적 최적화 방법과 비교하여 유리하다는 것을 보여줍니다. 마지막으로 무한대 규범에 기반한 *아담의* 변형인 *AdaMax에 대해 설명합니다.*

1 소개

확률적 기울기 기반 최적화는 많은 과학 및 공학 분야에서 실질적으로 중요한 핵심 요소입니다. 이러한 분야의 많은 문제는 매개변수에 대한 최대화 또는 최소화가 필요한 일부 스칼라 매개변수화된 목적 함수의 최적화로 캐스팅할 수 있습니다. 함수가 파라미터에 따라 미분 가능한 경우, 기울기 하강은 모든 파라미터에 대한 일차 편미분을 계산하는 것이 함수를 평가하는 것과 동일한 계산 복잡도를 갖기 때문에 비교적 효율적인 최적화 방법입니다. 종종 목적 함수는 확률적입니다. 예를 들어, 많은 목적 함수는 서로 다른 데이터 하위 샘플에서 평가된 하위 함수의 합으로 구성되며, 이 경우 점진적인 단계를 수행하여 최적화를 보다 효율적으로 수행할 수 있습니다.

개별 하위 함수, 즉 확률적 경사 하강(SGD) 또는 상승. SGD는 최근 딥러닝의 발전과 같은 많은 머신러닝 성공 사례에서 중심이 되는 효율적이고 효과적인 최적화 방법임이 입증되었습니다(Deng et al., 2013; Krizhevsky et al., 2012; Hinton & Salakhutdinov, 2006; Hinton et al., 2012a; Graves et al., 2013). 목표에는 데이터 서브샘플링 이외의 다른 노이즈 소스(예: 드롭아웃(Hinton et al., 2012b) 정규화)가 있을 수도 있습니다. 이러한 모든 노이즈가 있는 목표에 대해 효율적인 확률적 최적화 기법이 필요합니다. 이 백서의 초점은 고차원 매개변수 공간을 가진 확률적 목표의 최적화에 있습니다. 이

러한 경우 고차 최적화 방법은 적합하지 않으므로 본 백서에서는 일차 방법으로만 논의합니다.

메모리 요구량이 거의 없는 일차 그래디언트만 필요한 효율적인 확률론적 최적화 방법인 *Adam을* 제안합니다. 이 방법은 기울기의 첫 번째 및 두 번째 모멘트 추정으로부터 다양한 매개변수에 대한 개별 적응 학습률을 계산하며, Adam이라는 이름은 적응 모멘트 추정에서 유래했습니다. 이 방법은 최근에 널리 사용되는 두 가지 방법의 장점을 결합하도록 설계되었습니다: 희소 그래디언트에서 잘 작동하는 AdaGrad(Duchi et al., 2011)와 온라인 및 비고정 환경에서 잘 작동하는 RMSProp(Tieleman & Hinton, 2012); 이러한 방법과 다른 확률적 최적화 방법과의 중요한 연결은 섹션 5에 설명되어 있습니다. 아담의 장점은 파라미터 업데이트의 크기가 그라데이션의 크기 조정에 변하지 않고, 단계 크기가 단계 크기 하이퍼파라미터에 의해 대략적으로 제한되며, 고정된 목표가 필요하지 않고, 희박한 그라데이션에서 작동하며, 자연스럽게 일종의 단계 크기 어닐링을 수행한다는 점 등 여러 가지가 있습니다.

[*] 동등한 기여. Google 행아웃에서 동전 던지기로 저자 순서를 결정합니다.

알고리즘 1: 확률적 최적화를 위해 제안한 알고리즘인 *Adam.* 자세한 내용은 섹션 2를 참조하고, 조금 더 효율적인(그러나 덜 명확한) 계산 순서는 g^2 는 원소별 정사각형을 나타냅니다. $g_t g_t$. 테스트된 머신 러닝 문제에 적합한 기본 *설정은* $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ 입니다. 벡터에 대한 모든 연산은 요소 단위로 이루어집니다. β^t 및 β^t *에서* β_1 및 β_2 를 거듭제곱 t*로* 표시합니다.

필요: α: 단계 크기
Require: $\beta_1, \beta_2 \in [0, 1)$: 현재 추정치에 대한 지수 감쇠율입니다.
필요: $f(\theta)$: 파라미터 θ*가* 있는 확률적 목적 함수
필요: θ_0 : 초기 파라미터 벡터 m_0
$\rightarrow$ 0 (초기화 1st 모멘트 벡터) v_0
$\rightarrow$ 0 (초기화 2nd 모멘트 벡터) t
0 (타임스텝 초기화)
θ_t 수렴하지 않는 **동안**
$t \rightarrow t + 1$
$g_t \rightarrow \nabla f_{\theta t}(\theta_{t-1})$ (시간 간격 t*에서* 확률적 목표에 따른 기울기 구하기) m_t
$\rightarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (편향된 첫 순간 추정값 업데이트)
$v_t \rightarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (편향된 두 번째 원시 모멘트 추정치 업데이트)
$\hat{m}_t \rightarrow m_t / (1 - \beta_1)^t$ (편향 보정된 첫 순간 추정치 계산)
$\hat{v}_t \rightarrow v_t / (1 - \beta_2)^t$ (편향 보정된 두 번째 원시 모멘트 추정치 계산)
$\theta_t \rightarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$ (매개변수 업데이트)
while
반환 θ_t (결과 매개변수)

섹션 2에서는 알고리즘과 업데이트 규칙의 속성에 대해 설명합니다. 섹션 3에서는 초기화 편향 보정 기법을 설명하고, 섹션 4에서는 온라인 컨벡스 프로그래밍에서 아담의 수렴에 대한 이론적 분석을 제공합니다. 경험적으로, 섹션 6에서 볼 수 있듯이 다양한 모델과 데이터 세트에 대해 다른 방법보다 일관되게 우수한 성능을 보였습니다. 전반적으로, 우리는 Adam이 대규모 고차원 머신러닝 문제에 확장 가능한 다목적 알고리즘임을 보여줍니다.

2 알고리즘

우리가 제안한 알고리즘 *Adam의 의사* 코드는 알고리즘 1을 참조하세요. $f(\theta)$를 잡음이 있는 객체 함수, 즉 파라미터 θ*에 따라 미분* 가능한 확률적 스칼라 함수라고 가정합니다. 우리는 파라미터 θ*에 따라* 이 함수의 기대값인 $E[f(\theta)]$를 최소화하는 데 관심이 있습니다. $f_1(\theta), ..., f_T(\theta)$를 사용하면 후속 시간 간격 1, ..., T에서 확률 함수의 실현을 관찰할 수 있습니다. 확률성은 데이터 포인트의 무작위 하위 샘플(미니배치)에서의 평가에서 발생하거나 내재된 함수 노이즈에서 발생할 수 있습니다. $g_t = \nabla f_{\theta t}(\theta)$를 사용하면 기울기, 즉 시간 간격 t*에서* 평가된 f_t 의 부분 도함수 벡터를 나타낼 수 있습니다.

이 알고리즘은 기울기(m_t)와 제곱 기울기(v_t)의 지수 이동 평균을 업데이트하며, 하이퍼 파라미터 $\beta_1, \beta_2 \in [0, 1)$는 이러한 이동 평균의 지수 감쇠율을 제어합니다. 이동 평균 자체는 기울기의 1st 모멘트(평균)와 2nd 원시 모멘트(중심이 없는 분산)의 추정치입니다. 그러나 이러한 이동 평균은 0의 벡터로 초기화되므로 특히 초기 시간 간격 동안, 특히 감쇠율이 작을 때(즉, β*가* 1에 가까울 때) 모멘트 추정치

가 0에 편향되게 됩니다. 좋은 소식은 이러한 초기화 편향은 쉽게 대응할 수 있어 편향이 보정된 모멘트를 얻을 수 있다는 것입니다.
추정치 m_t 및 v . 자세한 내용은 섹션 3을 참조하세요.

알고리즘 1의 효율성은 명확성을 희생하더라도 계산 순서를 변경하여(예: 루프의 마지막 세 줄을 다음 줄로 대체) 개선할 수 있습니다:

$\alpha_t = \alpha - \quad 1 - \beta^t 2/(1 - \beta^t$) 및 $\vartheta_t \rightarrow \vartheta_{t-1} - \alpha_t - m_t /(\sqrt{v_t} + \epsilon$ ^).

2.1 아담의 업데이트 규칙

아담의 업데이트 규칙의 중요한 속성은 단계 크기를 신중하게 선택한다는 점입니다. $\epsilon = 0$이라고 가정하면 시간 간격 v에서 매개변수 공간에서 취한 유효 단계는 $\Delta_t = \alpha$ - m^ $_t$ /입니다. 유효 스텝 크기는 다음과 같습니다. 개의 상한이 있습니다: $|\Delta_t| \leq \alpha$ - (1 $\quad 1 - \beta_2$ 경우 $(1 - \beta_1) > > \quad 1 - \beta_2$, 및 $|\Delta_t| \leq \alpha$ $- \beta_1)$ / 입니다.

그렇지 않으면. 첫 번째 경우는 그라디언트의 희소성이 가장 심각한 경우에만 발생합니다.
는 현재 타임스텝을 제외한 모든 타임스텝에서 0 이 되어야 합니다. 희소성이 √보다 적은 경우, 유효 스텝 크기
는 더 작아집니다. (1 - β_1)== 1 - β_2 우리는 | m ^ $_t$/를 얻습니다. 따라서 $|\Delta_t| < \alpha$.
보다 일반적인 시나리오에서는 $|E[g]/\sqrt{E[g^2]}| \leq 1$. 효과적인
각 타임스텝에서 파라미터 공간에서 취한 스텝의 크기는 스텝 크기 설정 α, 즉 $\Delta_t \in \alpha$에 의해 대략적
으로 제한됩니다. 이는 현재 파라미터 값 주변에 *신뢰 영역을* 설정하는 것으로 이해할 수 있으며, 그
이상은 현재 기울기 추정치가 충분한 정보를 제공하지 못합니다. 이렇게 하면 일반적으로 α*의* 적절
한 스케일을 미리 알기가 비교적 쉽습니다. 예를 들어, 많은 머신 러닝 모델의 경우 매개변수 공간에
서 설정된 특정 영역 내에 좋은 최적값이 높은 확률로 존재한다는 것을 미리 알고 있는 경우가 많
으며, 매개변수에 대한 사전 분포가 있는 경우도 드물지 않습니다. α*는* 파라미터 공간에서 단계의
크기를 설정(상한)하기 때문에, 우리는 종종 옵티마가 다음과 같이 되도록 α*의* 올바른 크기 순서를 추
론할 수 있습니다.
에서 $\hat{v}$ w√까지 몇 번의 반복을 통해 도달할 수 있습니다. 약간의 용어 남용이 있습니다,
비율 $\hat{m}$)/√를 신호 *대 잡음비* (*SNR*)라고 합니다. SNR이 작을수록 효과적인
단계 크기 $_t\Delta$이 0에 가까워집니다. SNR이 작을수록 다음과 같은 장점이 있으므로 바람직한 속성
의 방향에 대한 업데이트.
그라데이션. 예를 들어, SNR 값은 일반적으로 최적값에 가까워질수록 0에 가까워져 매개변수 공간
에서 유효 스텝이 작아지는데, 이는 자동 어닐링의 한 형태입니다. 유효 스텝 크기 와 계수 c 및 v_t 는 상쇄됩니다: (c - m_t)/(

3 초기화 편향 보정

섹션 2에서 설명한 것처럼 아담은 초기화 편향 보정 항을 활용합니다. 여기서는 두 번째 모멘트
추정에 대한 항을 유도할 것이며, 첫 번째 모멘트 추정에 대한 유도법은 완전히 유사합니다. g*를* 확
률적 목표 f 의 기울기로 하고, 감쇠율 β_2 을 갖는 제곱 기울기의 지수 이동 평균을 사용하여 두
번째 원 모멘트(중심이 없는 분산)를 추정하고자 합니다. g_1 , ..., g_T 을 후속 시간 간격의 기울기
로 하고, 각각 기본 기울기 분포 g_t $p(g_t)$에서 끌어옵니다. 지수 이동 평균을 $v_0 = 0$(0의 벡터)으
로 초기화하겠습니다. 먼저 지수 이동 평균 $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g^2$ (여기서 g^2 는 원소 제
곱 $g_t \odot g_t$)를 나타내며, 다음과 같이 쓸 수 있습니다.
이전 모든 시간 간격의 그라디언트 함수입니다:

$$v_t = (1 - \beta_2) \sum_{i=1}^{t} \beta_2^{t-i} \cdot g_i^2 \tag{1}$$

시간 간격 t*에서* 지수 이동 평균의 기대값인 $E[v_t]$가 실제 두 번째 순간 $E[g^2]$와 어떻게 관련되는지
알고 싶으므로 둘 사이의 불일치를 보정할 수 있습니다. 방정식 (1)의 왼쪽과 오른쪽의 기대값을 구합
니다:

$$E[v_t] = E\left[(1 - \beta_2) \sum_{i=1}^{t} \beta_2^{t-i} \cdot g_i^2\right] \tag{2}$$

$$= \mathrm{E}[g_t^2] - (1 - \beta_2) \sum_{i=1}^{t} \beta_2^{t-i} + \zeta \tag{3}$$

$$= \mathrm{E}[g_t^2] - (1 - \beta_2^t) + \zeta \tag{4}$$

여기서 $\zeta = 0$은 실제 두 번째 모멘트 $\mathrm{E}[g^2]$가 교정된 경우이고, 그렇지 않은 경우 지수 감쇠율 β_1을 선택하면 지수 이동 평균이 과거에 너무 먼 기울기에 작은 가중치를 할당하도록 ζ를 작게 유지할 수 있습니다(그리고 그래야 합니다). 남은 것은 $(1 - \beta_2^t)$이며, 이는 이동 평균을 0으로 초기화하기 때문에 발생합니다. 따라서 알고리즘 1에서는 초기화 편향을 수정하기 위해 이 항으로 나눕니다.

희박한 기울기의 경우, 두 번째 모멘트를 안정적으로 추정하려면 β_2의 작은 값을 선택하여 많은 기울기에 대한 평균을 구해야 하지만, 초기화 편향 보정이 부족하면 초기 단계가 훨씬 커질 수 있는 β_2의 작은 경우와 정확히 일치하는 경우입니다.

4 컨버전스 분석

(Zinkevich, 2003)에서 제안한 온라인 학습 프레임워크를 사용하여 아담의 수렴을 분석합니다. 임의의 미지의 볼록 비용 함수 $f_1(\theta), f_2(\theta), \ldots, f_T(\theta)$의 임의의 시퀀스가 주어집니다. 각 시간 t에서 우리의 목표는 파라미터 θ_t를 예측하고 이전에 알려지지 않은 비용 함수 f_t에 대해 평가하는 것입니다. 시퀀스의 특성을 미리 알 수 없으므로, 이전 모든 단계에 대해 가능한 집합에서 온라인 예측 $f_t(\theta_t)$와 최상의 고정점 파라미터 $f_t(\theta^*)$ 사이의 이전 차이를 모두 합한 값인 후회(regret)를 사용하여 알고리즘을 평가합니다. 구체적으로 후회는 다음과 같이 정의됩니다:

$$R(T) = \sum_{t=1}^{T} [f_t(\theta_t) - f_t(\theta^*)] \tag{5}$$

여기서 $\theta^* = \arg\min_{\theta \in \mathcal{X}} \sum_{t=1}^{T} f_t(\theta)$. 아담이 $O(\sqrt{T})$ 후회 바인딩을 가지고 있음을 보여주고 그 증명이 주어집니다. 를 참조하세요. 우리의 결과는 이 일반적인 볼록 온라인 학습 문제에 대해 가장 잘 알려진 바운드와 비슷합니다. 또한 몇 가지 정의를 사용하여 표기법을 단순화했는데, 여기서 g_t, $f_t(\theta_t)$ 및 $g_{t,i}$을 i^{th} 요소로 정의합니다. $g_{1:t,i}$ $\mathbb{R}^t$을 그라디언트의 i^{th} 차원을 포함하는 벡터로 정의합니다. t까지 모든 반복에 대해 $g_{1:t,i} = [g_{1,i}, g_{2,i}, \ldots, g_{t,i}]$. 또한 γ, $\frac{\sqrt{1-\beta_2}}{1-\beta_2}$을 정의합니다. 우리의 다음 학습률 α_t이 t^{-2}의 속도로 감쇠하고 첫 번째 순간 실행 평균 계수 $\beta_{1,t}$가 λ에 따라 기하급수적으로 감쇠하는 경우(예: $1 - 10^{-8}$) 정리가 성립합니다.

정리 4.1. *함수 f_t가 $\|\nabla f_t(\theta)\|_2 \le G$, $\|\nabla f_t(\theta)\|_\infty \le G_\infty$인 경계 경사도를 가지고 있다고 가정합니다. 모든 $\theta \in \mathbb{R}^d$ 및 Adam이 생성한 모든 θ_t 사이의 거리는 $\|\theta_n - \theta_m\|_2 \le D$에 대해 $\|\theta_m - \theta_n\|_\infty \le D_\infty$ 모든 $m, n \in \{1, \ldots, T\}$, $\beta_1, \beta_2 \in [0, 1)$은 $\frac{\sqrt{1-\beta_2}}{1-\beta_2} < 1$을 만족합니다. $\alpha_t = \frac{\alpha}{\sqrt{t}}$라고 하자. $\beta_{1,t} = \beta_1 \lambda^{t-1}$, $\lambda \in (0, 1)$. 아담은 모든 $T \ge 1$에 대해 다음과 같은 보장을 얻습니다.*

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

정리 4.1은 데이터 특징이 희박하고 경계가 있는 그라데이션일 때, 합상한 $\sum_{i=1}^{d}$보다 훨씬 작을 수 있습니다. $\|g_{1:T,i}\|_2 << dG_\infty \sqrt{T}$ 및 $\sum_{i=1}^{d} \sqrt{T} \hat{v}_{T,i} << dG_\infty \sqrt{T}$ 특히 함수 및 데이터 기능의 클래스가 다음과 같은 형태인 경우 섹션 1.2를 참조하십시오(Duchi et al., 2011). 기대값 $E[\sum_{i=1}^{d} \|g_{1:T,i}\|_2]$ 또한 적용 를 아담에게 전달합니다. 특히 아담과 아다그라드와 같은 적응형 방식은 $O(\log d)$ 는 비적응적 방법의 $O(\sqrt{dT})$보다 개선되었습니다. $\beta_{1,t}$를 0으로 감소시키는 것은 이론적 분석에서 중요한 의미를 가지며, 훈련 종료 시 운동량 계수를 줄이면 컨버전스를 개선할 수 있다는 이전의 경험적 연구 결과와도 일치합니다(예: Sutskever et al., 2013).

마지막으로 아담의 평균 후회가 수렴하는 것을 보여줄 수 있습니다,

정리 4.2. *함수 f_t가 모든 $\theta \in \mathbb{R}^d$에 대해 $\|\nabla f_t(\theta)\|_2 \le G$, $\|\nabla f_t(\theta)\|_\infty \le G_\infty$의 경계 기울기를 가지며, 아담이 생성한 모든 θ_t 사이의 거리는 $\|\theta_n - \theta_m\|_2 \le D$, $\|\theta_m - \theta_n\|_\infty \le D_\infty$의 경계가 있다고 가정합니다. 아담은 모두에 대해 다음과 같은 보장을 달성합니다. $T \ge 1$.*

$$\frac{R(T)}{T} = O\left(\frac{1}{\sqrt{T}}\right)$$

이 결과는 정리 4.1과 $\sum_{i=1}^{d} \|g_{1:T,i}\|_2 \le dG_\infty \sqrt{T}$를 사용하여 얻을 수 있습니다. 따라서

$\lim_{T \to \infty} \frac{R(T)}{T} = 0$.

5 관련 작업

아담과 직접적인 관련이 있는 최적화 방법으로는 RMSProp(Tieleman & Hinton, 2012; Graves, 2013)과 AdaGrad(Duchi et al., 2011)가 있으며, 이러한 관계는 아래에서 설명합니다. 다른 확률적 최적화 방법으로는 vSGD(Schaul et al., 2012), AdaDelta(Zeiler, 2012), Roux & Fitzgibbon(2010)의 자연 뉴턴 방법 등이 있으며, 모두 곡률을 추정하여 단계화를 설정합니다.

을 1차 정보로부터 구합니다. 함수의 합 최적화(SFO)(Sohl-Dickstein 외, 2014)는 미니배치를 기반으로 하는 준 뉴턴 방법이지만, (Adam과 달리) 데이터 집합의 미니배치 파티션 수에 선형적인 메모리 요구 사항을 가지므로 GPU와 같은 메모리 제약이 있는 시스템에서는 종종 실행이 불가능합니다. 자연 경사 하강(NGD)과 마찬가지로(Amari, 1998), Adam은
v_t 는 대각선에 대한 근사치이므로 데이터의 기하학적 구조에 적응하는 전제 조건을 사용합니다.
의 대각선 피셔 정보 행렬의 역의 제곱근으로 전제 조건화하여 바닐라 NGD보다 더 보수적으로 적응하지만, 아담의 전제 조건화기(AdaGrad와 마찬가지로)는 대각선 피셔 정보 행렬 근사치의 역의 제곱근으로 전제 조건화하여 바닐라 NGD보다 적응이 더 보수적입니다.

RMSProp: 아담과 밀접한 관련이 있는 최적화 방법은 RMSProp입니다(Tieleman & Hinton, 2012). 모멘텀이 있는 버전이 사용되기도 합니다(Graves, 2013). 모멘텀이 있는 RMSProp과 Adam 사이에는 몇 가지 중요한 차이점이 있습니다. 모멘텀이 있는 RMSProp은 재조정된 기울기의 모멘텀을 사용하여 파라미터 업데이트를 생성하는 반면, Adam 업데이트는 기울기의 첫 번째 및 두 번째 모멘텀의 실행 평균을 사용하여 직접 추정합니다. 6.4절에서 경험적으로 증명한 바와 같이 β_2 값이 1에 가까운 경우(희박한 기울기의 경우 필요)에는 편향 보정 항이 없는데, 이 경우 편향을 보정하지 않으면 매우 큰 단계 크기와 종종 발산이 발생하기 때문에 이 점이 가장 중요합니다.

AdaGrad: 희박한 그라디언트에 잘 작동하는 알고리즘은 AdaGrad입니다(Duchi et al., 2011). 그것의 기본 버전은 매개변수를 $\vartheta_{t+1} = \vartheta_t - \alpha \cdot g_t / \sqrt{\sum_{i=1}^{t} g_t^2}$ 로 업데이트합니다. β_2 를 다음과 같이 선택하면 아래에서 1에 무한히 가까워지면, 제한 $\lim_{\beta_2 \to 1} \hat{v}_t = t^{-1} \cdot \sum_{i=1}^{t} g_t^2$. AdaGrad는 $\beta_1 = 0$, 무한소수 $(1 - \beta_2)$ 의 아담 버전과 α 를 어닐링 버전으로 대체합니다.
$\alpha_t = \alpha \cdot t^{-1/2}$, 즉 $\vartheta_t - \alpha \cdot t^{-1/2} \cdot \hat{m}_t / \sqrt{\lim_{\beta_2 \to 1} \hat{v}_t} = \vartheta_t - \alpha \cdot t^{-1/2} \cdot g_t / \sqrt{t^{-1} \cdot \sum_{i=1}^{t} g_t^2} = \vartheta_t - \alpha \cdot g_t / \sqrt{\sum_{i=1}^{t} g_t^2}$. 바이어스 보정 항을 제거할 때 유지되지 않으며, RMSProp에서와 같이 ... β_2 를 1에 무한히 가깝게 설정하면 편향이 무한히 커지고 파라미터 업데이트도 무한히 커집니다.

6 실험

제안한 방법을 실증적으로 평가하기 위해 로지스틱 회귀, 다층 완전 연결 신경망, 심층 합성곱 신경망 등 널리 사용되는 다양한 머신러닝 모델을 조사했습니다. 대규모 모델과 데이터 세트를 사용하여 Adam이 실제 딥러닝 문제를 효율적으로 해결할 수 있음을 입증했습니다.

서로 다른 최적화 알고리즘을 비교할 때는 동일한 파라미터 초기화를 사용합니다. 학습 속도 및 모멘텀과 같은 하이퍼 파라미터는 조밀한 그리드에서 검색되며 최적의 하이퍼 파라미터 설정을 사용하여 결과가 보고됩니다.

6.1 실험: 로지스틱 회귀

우리는 MNIST 데이터 세트를 사용하여 L2 정규화된 다중 클래스 로지스틱 회귀에 대해 제안한 방법을 평가합니다. 로지스틱 회귀는 잘 연구된 볼록한 목표를 가지고 있어 비교에 적합합니다.
로컬 최소 문제에 대해 √ 걱정할 필요 없이 다양한 옵티마이저를 사용할 수 있습니다. 로지스틱스의 단계 크기 α 는
회귀 실험은 *1/t* 붕괴, 즉 $\alpha_t = \frac{\alpha}{\sqrt{t}}$ 로 조정되어 우리의 이론과 일치합니다.

섹션 4의 통계적 예측을 사용합니다. 로지스틱 회귀는 784개의 차원 이미지 벡터에서 직접 클래스 라벨을 분류합니다. 128의 미니배치 크기를 사용하여 아담과 가속화된 SGD, 네스테로프 모멘텀 및 아다그라드를 비교합니다. 그림 1에 따르면, 아담은 모멘텀이 있는 SGD와 유사한 수렴을 산출하며, 둘 다 아다그라드보다 더 빠르게 수렴하는 것을 알 수 있습니다.

(Duchi et al., 2011)에서 논의한 바와 같이, Adagrad는 희박한 특징과 그라데이션을 효율적으로 처리할 수 있습니다.
를 주요 이론적 결과 중 하나로 삼는 반면, SGD는 희귀 기능 학습에 취약합니다. 아담과
1/t 감쇠의 스텝 사이즈는 이론적으로 Adagrad의 성능과 일치해야 합니다. 우리는

(Maas et al., 2011)의 IMDB 영화 리뷰 데이터 세트를 사용하여 스파스 특징 문제를 해결합니다. 우리는 IMDB 영화 리뷰를 가장 빈번한 첫 10,000개의 단어를 포함하는 단어 가방(BoW) 특징 벡터로 사전 처리합니다. 각 리뷰에 대한 10,000차원의 BoW 특징 벡터는 매우 희박합니다. (Wang & Manning, 2013)에서 제안한 바와 같이, 50%의 드롭아웃 노이즈가 다음과 같은 동안 BoW 특징에 적용될 수 있습니다.

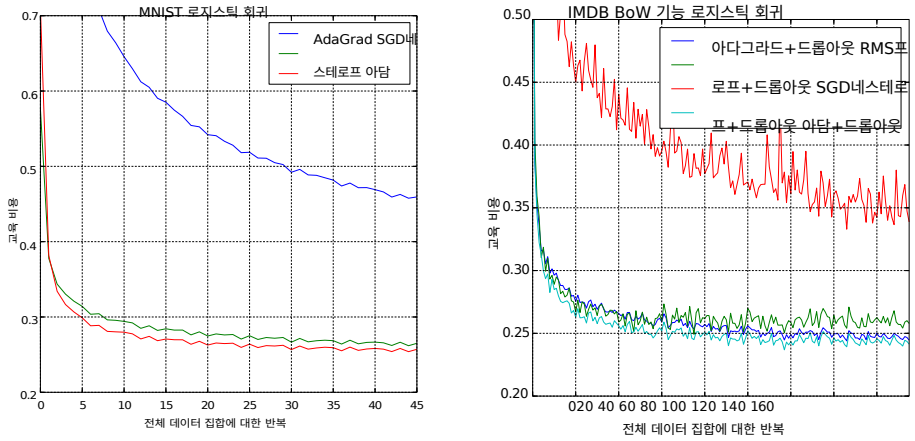


그림 1: 10,000개의 단어 가방(BoW) 특징 벡터를 사용하여 MNIST 이미지와 IMDB 영화 리뷰에 대한 음의 로그 가능성을 로지스틱 회귀 학습합니다.

트레이닝을 통해 과도한 피팅을 방지합니다. 그림 1에서 Adagrad는 드롭아웃 노이즈가 있을 때와 없을 때 모두 Nesterov 모멘텀을 사용하는 SGD보다 큰 폭으로 성능이 뛰어납니다. 아담은 아다그라드만큼 빠르게 수렴합니다. 아담의 경험적 성능은 섹션 2와 4의 이론적 결과와 일치합니다. 아담은 아다그라드와 유사하게 희소 특징을 활용하고 모멘텀이 있는 일반 SGD보다 더 빠른 수렴 속도를 얻을 수 있습니다.

6.2 실험: 다층 신경망

다층 신경망은 비볼록 목적 함수를 가진 강력한 모델입니다. 수렴 분석은 비볼록 문제에는 적용되지 않지만, 경험적으로 이러한 경우 아담이 다른 방법보다 더 나은 성능을 발휘하는 경우가 많다는 것을 발견했습니다. 실험에서는 이 분야의 이전 논문과 일치하는 모델을 선택했으며, 각각 1000개의 숨겨진 유닛이 있는 완전히 연결된 두 개의 숨겨진 레이어와 ReLU 활성화가 있는 신경망 모델을 128개의 미니 배치 크기로 실험에 사용했습니다.

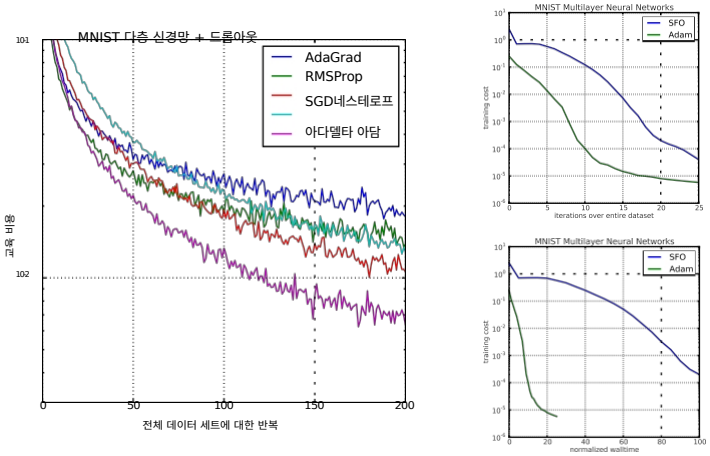
먼저, 과적합을 방지하기 위해 매개변수에 대한 L_2 가중치 감쇠가 있는 표준 결정론적 교차 엔트로피 목적 함수를 사용하여 다양한 최적화 기법을 연구합니다. 함수의 합(SFO) 방법(Sohl-Dickstein 외., 2014)은 최근에 제안된 준 뉴턴 방법으로, 데이터의 미니배치에서 작동하며 다층 신경망의 최적화에서 좋은 성능을 보였습니다. 이러한 모델을 훈련하기 위해 이를 구현하고 Adam과 비교했습니다. 그림 2는 반복 횟수와 월 클럭 시간 모두에서 Adam이 더 빠르게 학습하는 것을 보여줍니다. 곡률 정보를 업데이트하는 데 드는 비용으로 인해 SFO는 Adam에 비해 반복 횟수당 5~10배 느리고, 메모리 요구량은 미니배치 수에 따라 선형적으로 증가합니다.

드롭아웃과 같은 확률 정규화 방법은 과적합을 방지하는 효과적인 방법이며, 그 단순성 때문에 실무에서 자주 사용됩니다. SFO는 결정론적 하위 함수를 가정하며, 실제로 확률적 정규화를 통해 비용 함수에 수렴하는 데 실패했습니다. 드롭아웃 노이즈로 훈련된 다층 신경망에서 다른 확률론적 1차 방법

과 Adam의 효과를 비교해 보았습니다. 그림 2는 그 결과를 보여주는데, 아담은 다른 방법보다 더 나은 수렴을 보여줍니다.

6.3 실험하기: 컨볼루션 신경망

여러 계층의 컨볼루션, 풀링 및 비선형 유닛으로 구성된 컨볼루션 신경망(CNN)은 컴퓨터 비전 작업에서 상당한 성공을 거두었습니다. 대부분의 완전히 연결된 신경망과 달리 CNN의 가중치 공유는 레이어마다 매우 다른 그라데이션을 만들어냅니다. 실제로 SGD를 적용할 때는 컨볼루션 레이어의 학습 속도가 더 작은 경우가 많습니다. 심층 CNN에서 아담의 효과를 보여드립니다. Facebook의 CNN 아키텍처는 5x5 컨볼루션 필터와 3x3 최대 풀링(보폭 2)의 세 단계를 번갈아 가며 사용하고, 그 다음에는 1000개의 정류된 선형 숨겨진 유닛(ReLU)으로 완전히 연결된 레이어를 사용합니다. 입력 이미지는 화이트닝을 통해 사전 처리됩니다.



(a) (b)

그림 2: MNIST 이미지에 대한 다층 신경망 훈련. (a) 드롭아웃 확률 정규화를 사용한 신경망. (b) 결정론적 비용 함수를 사용하는 신경망. 함수의 합(SFO) 최적화 기법과 비교합니다(Sohl-Dickstein 외 ., 2014).

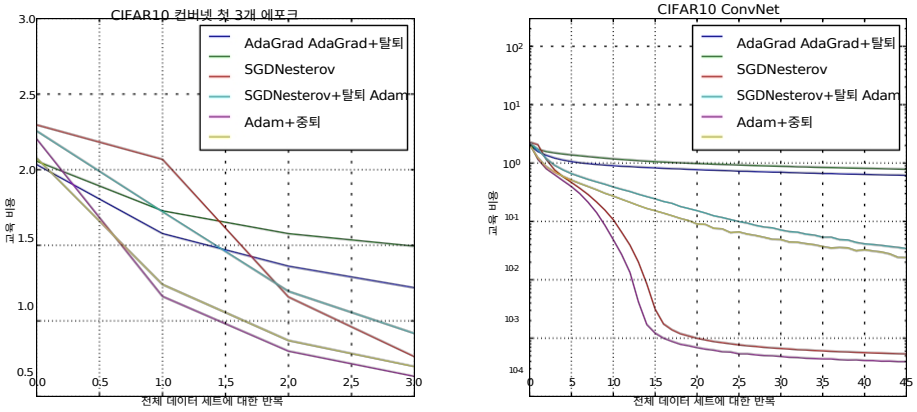
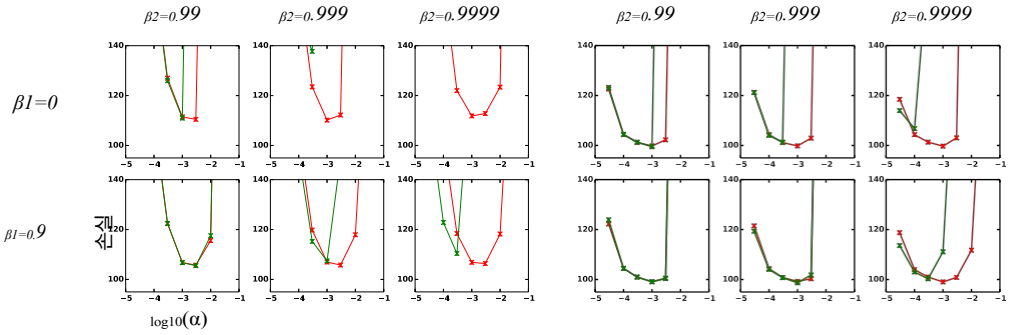


그림 3: 컨볼루션 신경망 훈련 비용. (왼쪽) 처음 세 번의 에포크에 대한 훈련 비용. (오른쪽) 45개 에포크에 대한 훈련 비용. c64-c64-c128-1000 아키텍처의 CIFAR-10.

드롭아웃 노이즈가 입력 레이어와 완전히 연결된 레이어에 적용됩니다. 미니배치 크기도 이전 실험과 유사하게 128로 설정합니다.

흥미롭게도 그림 3(왼쪽)과 같이 훈련 초기 단계에서는 아담과 아다그라드 모두 비용을 낮추며 빠르게 발전하지만, 결국 그림 3(오른쪽)에 표시된 CNN의 경우 아담과 SGD가 아다그라드보다 훨씬 더 빠르게 수렴합니다. 두 번째 모멘트 추정치 v_t 가 몇 번의 에포크 후에 0으로 사라지고 알고리즘 1의 ϵ $/\!/$ 의해 지배되는 것을 알 수 있습니다. 따라서 두 번째 모멘트 추정치는 6.2절의 완전 연결 네트워크와 비교할 때 CNN에서 비용 함수의 기하학적 구조에 대한 근사치가 좋지 않습니다. 반면, 첫 번째 모

멘트를 통해 미니 배치 분산을 줄이는 것이 CNN에서 더 중요하며 속도 향상에 기여합니다. 결과적으로 이 특정 실험에서 Adagrad는 다른 네트워크보다 훨씬 느리게 수렴합니다. 아담은 모멘텀을 통해 SGD에 비해 미미한 개선을 보이지만, SGD에서처럼 수동으로 선택하는 대신 학습 속도 척도를 다른 레이어에 맞게 조정합니다.



(a) 10회 이후 (b) 100회 후

그림 4: 단계 크기 α(x축)와 하이퍼 파라미터 β_1 및 β_2 의 다양한 설정에 대해 10개의 에포크(왼쪽) 및 100개의 에포크(오른쪽) 후 바이어스 보정 조건(빨간색 선)과 바이어스 보정 조건 없음(녹색 선)이 손실(y축)에 미치는 영향(Kingma & Welling, 2013), 변형 자동 인코더(VAE) 학습 시.

6.4 실험: 편향 보정 용어

또한 섹션 2와 3에서 설명한 편향성 보정 용어의 효과를 경험적으로 평가합니다. 섹션 5에서 설명한 바와 같이 편향 보정 조건을 제거하면 운동량을 가진 RMSProp(Tiele- man & Hinton, 2012) 버전이 생성됩니다. 소프트플러스 비선형성을 가진 500개의 숨겨진 유닛과 50차원 구형 가우시안 잠재 변수가 있는 단일 숨겨진 레이어로 (Kingma & Welling, 2013)과 동일한 아키텍처의 가변 자동 인코더 (VAE)를 훈련할 때 β_1 및 β_2 을 변경합니다. β_1 [0, 0.9] 및 β_2 [0.99, 0.999, 0.9999], $\log_{10}(\alpha)$ [5, ..., 1] 과 같은 광범위한 하이퍼파라미터 선택에 대해 반복했습니다. 희박한 기울기에 대한 견고성에 필요한 β_2 값이 1에 가까울수록 초기화 편향이 커지므로, 편향 보정 항은 이러한 느린 감쇠의 경우 최적화에 부정적인 영향을 미치지 않도록 하는 데 중요할 것으로 예상됩니다.

그림 4에서 1에 가까운 β_2 값은 편향 보정 항이 없을 때, 특히 훈련의 처음 몇 회기에 훈련의 불안정성을 초래합니다. (1 β_2)의 작은 값과 편향 보정을 사용하면 가장 좋은 결과를 얻을 수 있었으며, 이는 숨겨진 단위가 특정 패턴에 특화됨에 따라 기울기가 더 희박해지는 경향이 있는 최적화 후반으로 갈수록 더욱 분명해졌습니다. 요약하면, Adam은 하이퍼 파라미터 설정에 관계없이 RMSProp과 동등하거나 더 나은 성능을 보였습니다.

7 확장 기능

7.1 ADAMAX

Adam에서 개별 가중치에 대한 업데이트 규칙은 개별 현재 및 과거 기울기의 (스케일링된) L^2 규범에 반비례하여 기울기를 스케일링하는 것입니다. L^2 규범 기반 업데이트 규칙을 L^p 규범 기반 업데이

트 규칙으로 일반화할 수 있습니다. 이러한 변형은 대규모의 경우 수치적으로 불안정해집니다. p. 그러나 p를 허용하는 $\rightarrow \infty$ 특별한 경우에는 놀랍도록 간단하고 안정적인 알고리즘이 등장합니다(알고리즘 2 참조). 이제 알고리즘을 도출해 보겠습니다. L^p 규범의 경우, 단계 크기

는 시간 t에서 $v^{1/p}$ 에 반비례합니다: $_t$

$$v_t = \beta_2^p v_{t-1} + (1 - \beta_2^p)|g_t|^p \tag{6}$$

$$= (1 - \beta_2^p) \sum_{i=1}^{t} \beta_2^{p(t-i)} \cdot |g_i|^p \tag{7}$$

알고리즘 2: 무한대 규범에 기반한 아담의 변형인 *AdaMax*. 자세한 내용은 섹션 7.1을 참조하세요. 테스트된 머신 러닝 문제에 대한 좋은 기본 설정은 $\alpha = 0.002$, $\beta_1 = 0.9$, $\beta_2 = 0.999$입니다. 여기서 $(\alpha/(1 - \beta_1^t))$는 β_1^t 를 β_1 를 거듭제곱한 학습률입니다.
편향 보정 용어가 필요합니다. 벡터에 대한 모든 연산은 요소 단위로 이루어집니다.

필요: α: 단계 크기
Require: $\beta_1, \beta_2 \in [0, 1)$: 지수 감쇠율
필요: $f(\theta)$: 파라미터 θ가 있는 확률적 목적 함수
필요: θ_0 : 초기 파라미터 벡터
$m_0 \leftarrow 0$ (1st 모멘트 벡터 초기화)
$u_0 \leftarrow 0$ (지수 가중 무한대 노름 초기화)
$t \leftarrow 0$ (타임스텝 초기화)
θ_t 수렴하지 않는 **동안**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla f_{\theta_t}(\theta_{t-1})$ (시간 간격 t에서 확률적 목표에 따른 기울기 구하기)
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (편향된 첫 순간 추정값 업데이트)
$u_t \leftarrow \max(\beta_2 \cdot u_{t-1}, |g_t|)$ (지수 가중 무한대 노름 업데이트)
$\theta_t \leftarrow \theta_{t-1} - (\alpha/(1 - \beta_1^t)) \cdot m_t/u_t$ (파라미터 업데이트)
while
반환 θ_t (결과 매개변수)

여기서 감쇠 항은 β_2 대신 β^p 로 동등하게 파라미터화되어 있습니다. 이제 $p \to \infty$ 가 됩니다,
를 입력한 다음 $u_t = \lim_{p\to\infty} (v)_t^{1/p}$
를 정의합니다:

$$(1 - \beta^p) \sum_{i=1}^{t} \beta^{p(t-i)} \cdot |g_i|^p \Big)^{1/p} \tag{8}$$

$$u_t = \lim_{p\to\infty} (v)_t^{1/p} = \lim_{p\to\infty} \Big((1 - \beta^p) \sum_{i=1}^{t} \beta^{p(t-i)} \cdot |g_i|^p \Big)^{1/p} \tag{9}$$

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

$$= \text{최대} \; \beta^{t-1} |g_1|, \beta^{t-2} |g_2|, \ldots, \beta |g_{t-1}|, |g_t| \tag{11}$$

이는 놀랍도록 간단한 재귀 공식에 해당합니다:

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

초기 값 $u_0 = 0$입니다. 이 경우 편리하게도 초기화 편향을 수정할 필요가 없다는 점에 유의하세요. 또한 매개변수 업데이트의 크기는 아담보다 AdaMax에서 더 간단한 바인딩을 가집니다: $|\Delta_t| \leq \alpha$.

7.2 시간 평균

마지막 반복은 확률적 근사치로 인해 노이즈가 발생하기 때문에 평균화를 통해 더 나은 일반화 성능을 얻을 수 있습니다. 이전에 Moulines & Bach(2011)에서 Polyak-Ruppert 평균화(Polyak & Juditsky, 1992; Ruppert, 1988)는 표준값의 수렴을 개선하는 것으로 나타났습니다.
SGD 여기서 $\bar{\theta}_t = \frac{1}{n} \sum_{i=1}^{n} \theta_i$ θ_t 또는 매개 변수에 대한 지수 이동 평균을 사용할 수도 있습니다.
를 사용하여 최신 파라미터 값에 더 높은 가중치를 부여합니다. 이는 간단하게 구현할 수 있습니다.
알고리즘 1과 2의 내부 루프에 한 줄을 추가하여 $\bar{\theta}_t \leftarrow \beta_2 \cdot \bar{\theta}_{t-1} + (1 - \beta_2)\theta_t$, $\bar{\theta}_0 = 0$입니다. 초기화 편향은 다시 추정기 $\hat{\theta}_t = \bar{\theta}_t/(1 - \beta_2^t)$로 보정할 수 있습니다.

8 결론

우리는 확률적 목적 함수의 그래디언트 기반 최적화를 위한 간단하고 계산적으로 효율적인 알고리즘을 도입했습니다. 이 방법은 다음과 같은 머신 러닝 문제를 대상으로 합니다.

대규모 데이터 세트 및/또는 고차원 매개변수 공간에 적합합니다. 이 방법은 최근 널리 사용되는 두 가지 최적화 방법, 즉 희박한 경사도를 처리하는 AdaGrad의 기능과 비고정 목표를 처리하는 RMSProp의 기능을 결합한 것입니다. 이 방법은 구현이 간단하고 메모리도 거의 필요하지 않습니다. 실험을 통해 볼록한 문제에서의 컨버전스 비율에 대한 분석을 확인했습니다. 전반적으로 아담은 강력하고 현장 머신러닝의 다양한 비볼록 최적화 문제에 적합하다는 것을 알 수 있었습니다.

9 감사

이 논문은 구글 딥마인드의 지원이 없었다면 존재하지 못했을 것입니다. 이보 다니헬카와 아담이라는 이름을 지어준 톰 솔에게 특별한 감사를 표합니다. 원래 AdaMax 파생에서 오류를 발견해준 듀크 대학교의 카이 팬(Kai Fan)에게도 감사드립니다. 이 작업의 실험은 부분적으로 네덜란드 국립 e-인프라에서 SURF 재단의 지원으로 수행되었습니다. 디에릭 킹마는 딥 러닝 분야의 구글 유럽 박사 펠로우십의 지원을 받았습니다.

참조

아마리, 슌이치. 자연 그라데이션은 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, He, Xiaodong, Williams, Jason 외. 최근 Microsoft에서 음성 연구를 위한 딥 러닝의 발전. *ICASSP 2013*, 2013.

두치, 존, 하잔, 엘라드, 싱어, 요람. 온라인 학습 및 확률적 최적화를 위한 적응형 하위 경사 방법. *기계 학습 연구 저널*, 12:2121-2159, 2011.

그레이브스, 알렉스. 반복 신경망을 이용한 시퀀스 생성. *arXiv preprint arXiv:1308.0850*, 2013.

그레이브스, 알렉스, 모하메드, 압델-라흐만, 힌튼, 제프리. 심층 순환 신경망을 사용한 음성 인식. *음향, 음성 및 신호 처리(ICASSP), 2013 IEEE 국제 컨퍼런스에서*,
pp. 6645-6649. IEEE, 2013.

힌튼, G.E. 및 살라쿠트디노프, R.R. 신경망으로 데이터의 차원 축소. *Science*, 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 외. 음성 인식에서 음향 모델링을 위한 심층 신경망: 네 연구 그룹의 공통된 견해. *신호 처리 매거진, IEEE*, 29(6):82-97, 2012a.

힌튼, 제프리 E, 스리바스타바, 니티쉬, 크리제프스키, 알렉스, 수츠케버, 일리야, 살라쿠트디노프, 루슬란 R. 특징 검출기의 공동 적응을 방지하여 신경망을 증명하는 방법. *arXiv preprint arXiv:1207.0580*, 2012b.

킹마, 디데릭 P 및 웰링, 맥스. 자동 인코딩 변형 베이즈. *제2회 학습 표현에 관한 국제 컨퍼런스(ICLR)*, 2013.

크리제프스키, 알렉스, 수츠케버, 일리야, 힌튼, 제프리 E. 심층 컨볼루션 신경망을 사용한 이미지넷 분류. *신경 정보 처리 시스템의 발전*, 1097-1105쪽, 2012.

Maas, Andrew L, Daly, Raymond E, Pham, Peter T, Huang, Dan, Ng, Andrew Y 및 Potts, Christopher. 감정 분

석을 위한 단어 벡터 학습. *제49회 전산 언어학 협회 연례 회의 논문집: 인간 언어 기술-제1 권*, 142-150쪽. 전산 언어학 협회, 2011.

물린스, 에릭 및 바흐, 프란시스 R. 기계 학습을 위한 확률적 근사 알고리즘의 비점근 분석. *신경 정보 처리 시스템의 발전*, 451-459쪽, 2011.

파스카누, 라즈반 및 벤지오, 요슈아. 심층 네트워크를 위한 자연 그라데이션의 재검토. *arXiv 프리프린트 arXiv:1301.3584*, 2013.

Polyak, Boris T 및 Juditsky, Anatoli B. 평균을 통한 확률적 근사치의 가속화. *SIAM 제어 및 최적화 저널*, 30(4):838-855, 1992.

루, 니콜라스 L 및 피츠기번, 앤드류 W. 빠른 자연 뉴턴 방법. *제27회 국제 기계 학습 컨퍼런스(ICML-10) 논문집,* 623-630쪽, 2010.

루퍼트, 데이비드. 천천히 수렴하는 로빈스-몬로 프로세스를 통한 효율적인 추정. 기술 보고서, 코넬 대학교 운영 연구 및 산업 공학, 1988.

숄, 톰, 장, 식신, 및 르쿤, 얀. 더 이상 성가신 학습률은 없습니다. *arXiv 사전 인쇄물 arXiv:1206.1106*, 2012.

Sohl-Dickstein, Jascha, 풀, 벤, 및 강굴리, 수리아. 스토캐스틱 그라데이션과 준 뉴턴 방법을 통합하여 대규모 최적화를 빠르게 수행합니다. *제31회 국제 기계 학습 컨퍼런스(ICML-14) 논문집,* 604-612쪽, 2014.

수츠케버, 일리야, 마르텐스, 제임스, 달, 조지, 힌튼, 제프리. 딥러닝에서 초기화와 모멘텀의 중요성. *제30회 국제 기계 학습 컨퍼런스(ICML-13) 논문집,* 1139-1147쪽, 2013.

강의 6.5 - RMSProp, COURSERA: 기계 학습을 위한 신경망, Tieleman, T. 및 Hinton, G. 강의 6.5. 기술 보고서, 2012.

왕, 시다 및 매닝, 크리스토퍼. 빠른 드롭아웃 훈련. *제30회 기계 학습 국제 컨퍼런스(ICML-13) 논문집,* 118-126쪽, 2013.

자일러, 매튜 D. 아다 델타: 적응형 학습 속도 방법. *아카이브 프리프린트 arXiv:1212.5701*, 2012. 징케비치, 마틴.

온라인 볼록 프로그래밍 및 일반화 된 무한 최소 경사 상승. 2003.

10 부록

10.1 컨버전스 증명

정의 10.1. *함수* $f : R^d \to R$ *은 모든* $x, y \in R^d$, *모든* $\lambda \in [0, 1]$에 *대해 볼록합니다,*

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

또한 볼록 함수는 접선의 하이퍼플레인에 의해 하한이 낮아질 수 있다는 점에 유의하세요.

정리 10.2. *함수* $f : R^d \to R$ *이 볼록하다면, 모든* $x, y \in R^d$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

위의 정리는 후회의 상한을 설정하는 데 사용할 수 있으며, 주요 정리에 대한 증명은 하이퍼플레인을 아담 업데이트 규칙으로 대체하여 구성할 수 있습니다.

다음 두 가지 정리는 우리의 주요 정리를 뒷받침하는 데 사용됩니다. 또한 표기법을 단순화하기 위해 몇 가지 정의를 사용하는데, 여기서 g_t , $\nabla f_t(\vartheta_t)$와 $g_{t,i}$ *을* i^{th} 요소로 정의합니다. $g_{1:t,i} \in \mathrm{R}^t$ *을* t*까지* 모든 반복에 걸쳐 기울기의 i^{th} 차원을 포함하는 벡터로 정의하면, $g_{1:t,i} = [g_{1,i}, g_{2,i}, \ldots, g_{t,i}]$ **정리 10.3.** $g_t = \nabla f_t(\vartheta_t)$ *및* $g_{1:t}$ *를 위와 같이 정의하고 경계가 지정되면,* $\|g_t\|_2 \leq G$, $\|g_t\|_\infty \leq G_\infty$,

$$\sum_{t=1}^{T} \frac{\hat{g}_{t,i}}{t} \leq 2G_\infty \|g_{1:T,i}\|_2$$

증명. T에 대한 귀납법을 사용해 부등식을 증명하겠습니다.

$T = 1$의 기본 사례는 다음과 같습니다. $\hat{g}_{1,i}^2 \leq 2G_\infty \|g_{1,i}\|_2$.

귀납적 단계의 경우,

$$\sum_{t=1}^{T} \frac{\hat{g}_{t,i}}{t} = \sum_{t=1}^{T-1} \frac{\hat{g}_{t,i}}{t} + \frac{\hat{g}_{T,i}}{T}$$

$$\leq 2G_\infty \|g_{1:T-1,i}\|_2 + \frac{\hat{g}_{T,i}}{T}$$

$$= 2G_\infty \sqrt{\|g_{1:T,i}\|_2^2 - g_{T,i}^2} + \frac{\hat{g}_{T,i}}{T}$$

보낸 사람, $\|g_{1:T,i}\|_2^2 - g_{T,i}^2 + \frac{g_{T,i}^4}{4\|g_{1:T,i}\|^2} \geq \|g_{1:T,i}\|_2 - \frac{g_{T,i}^2}{2\|g_{1:T,i}\|_2}$를 사용하면 양쪽 변의 제곱근과 가지고 있습니다,

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

$\leq \|g_{1:T,i}\|_2 - 2\sqrt{T}G_2 \qquad \frac{}{\infty}$

부등식을 재정렬하고 q를 대체합니다. $\overline{\|g_{1:T,i}\|_2^2 - g_{T,i}^2}$ 기간입니다,

$$G_\infty \frac{q}{\|g_{1:T,i}\|_2^2 - g_T^2} \leq \frac{s}{g_{T,i}^2} \leq 2G_\infty \|g_{1:T,i}\|_2$$

□

정리 10.4. *$\gamma, \sqrt{\beta_2}$을 보자. $\beta_1, \beta_2 \in [0, 1)$에 대해 $\frac{\beta_1}{\sqrt{\beta_2}} < 1$을 만족하고 경계가 있는 g_t, $\|g_t\|_2 \le G$, $\|g_t\|_\infty \le G_\infty$, 다음과 같은 부등식이 성립합니다.*

$$\sum_{t=1}^{T} \frac{m_{t,i}}{\sqrt{v_{t,i}}} \le \frac{2}{1-\gamma} \frac{1}{\sqrt{1-\beta_2}} \|g_{1:T,i}\|_2$$

증명. 가정 하에서 $\sqrt{1-\beta_t} \le \frac{1}{(1-\beta_1)^2}$. 합계에서 마지막 학기를 확장할 수 있습니다. 알고리즘 1의 업데이트 규칙을 사용합니다,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{m_{t,i}}{\sqrt{v_{t,i}}} &= \sum_{t=1}^{T-1} \frac{m_{t,i}}{\sqrt{v_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \sqrt{\frac{\sum_{k=1}^{T} (1-\beta_1)\beta_1^{T-k} g_{k,i})^2}{\sum_{j=1}^{T} (1-\beta_2)\beta_2^{T-j} g_{j,i}^2}} \\
&\le \sum_{t=1}^{T-1} \frac{m_{t,i}}{\sqrt{v_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \sum_{k=1}^{T} \sqrt{\frac{(1-\beta_1)\beta_1^{T-k} g_{k,i})^2}{(1-\beta_2)\beta_2^{T-k} g_{k,i}^2}} \\
&\le \sum_{t=1}^{T-1} \frac{m_{t,i}}{\sqrt{v_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \sum_{k=1}^{T} \frac{(1-\beta_1)\beta_1^{T-k}}{\sqrt{(1-\beta_2)\beta_2^{T-k}}} \|g_{k,i}\|_2 \\
&\le \sum_{t=1}^{T-1} \frac{m_{t,i}}{\sqrt{v_{t,i}}} + \frac{T}{\sqrt{T(1-\beta_2^T)}} \sum_{k=1}^{T} \gamma^{T-k} \|g_{k,i}\|_2
\end{aligned}$$

마찬가지로 합계에서 나머지 용어의 상한을 설정할 수 있습니다.

$$\begin{aligned}
\sum_{t=1}^{T} \frac{m_{t,i}}{\sqrt{v_{t,i}}} &\le \sum_{t=1}^{T} \frac{\|g_{t,i}\|_2}{\sqrt{t(1-\beta_2)}} \sum_{j=0}^{T-t} \gamma^j \\
&\le \sum_{t=1}^{T} \frac{\|g_{t,i}\|_2}{\sqrt{t(1-\beta_2)}} \sum_{j=0}^{T} \gamma^j
\end{aligned}$$

$\gamma < 1$의 경우 산술-기하 급수의 상한을 사용합니다, $\sum_{j=0}^{T} \gamma^j < \frac{1}{(1-\gamma)^2}$:

$$\sum_{t=1}^{T} \sum_{j=0}^{T} \gamma^j \le \frac{1}{(1-\gamma)2\sqrt{1-\beta_2}} \sum_{t=1}^{T} \frac{\|g_{t,i}\|_2}{\sqrt{t}}$$

정리 10.3을 적용합 $\frac{\|g_{t,i}\|_2}{\sqrt{t(1-\beta_2)}}$

니다,

$$\sum_{t=1}^{T} \frac{m_{t,i}}{\sqrt{v_{t,i}}} \le \frac{2G_\infty}{(1-\gamma)2\sqrt{1-\beta_2}} \| g_{1:T,i}\|_2$$

□

표기를 단순화하기 위해 γ 를 정의합니다, $\frac{\sqrt{\beta_2}}{\beta_1}$ 직관적으로 다음 정리는 다음과 같은 경우에 유효합니다.

학습률 α_t 은 t^{-2} 의 속도로 감쇠하고, 첫 번째 순간 실행 평균 계수 $\beta_{1,t}$ 는 λ 에 따라 기하급수적으로 감쇠하며, 일반적으로 1에 가까워집니다(예: $1 - 10^{-8}$).

정리 10.5. *함수 f_t 가 $\|\nabla f_t(\theta)\|_2 \leq G$, $\|\nabla f_t(\theta)\|_\infty \leq$ 인 경계 경사도를 갖는다고 가정합니다. 모든 $\theta \in R^d$ 및 Adam 이 생성한 모든 θ_t 사이의 거리는 $\theta_n - \theta_{m2} \leq D$ 에 대해 G_∞ ,*

$\|\vartheta_m - \vartheta_n\|_\infty \le D_\infty$ *모든* $m, n \in \{1, \ldots, T\}$, $\beta_1, \beta_2 \in [0, 1)$*은* $\frac{\beta_1}{\sqrt{\beta_2}} < 1$*을 만족합니다.* $\alpha_t = \frac{\alpha}{\sqrt{t}}$*라고 하자.* $\beta_{1,t} = \beta_1 \lambda^{t-1}$, $\lambda \in (0, 1)$. *아담은 모든* $T \ge 1$*에 대해 다음과 같은 보장을 얻습니다.*

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

증명. 정리 10.2를 사용하여 증명했습니다,

$$f_t(\vartheta_t) - f_t(\vartheta^*) \le g_t^T(\vartheta_t - \vartheta^*) = \sum_{i=1}^{d} g_{t,i}(\vartheta_{t,i} - \vartheta_i^*)$$

알고리즘 1에 제시된 업데이트 규칙에서,

$$\vartheta_{t+1} = \vartheta_t - \alpha_t \hat{m}_t / \sqrt{\hat{v}_t}$$
$$= \vartheta_t - \frac{\alpha_t}{1-\beta_1^t} \left(\frac{\beta_{1,t}}{\sqrt{\hat{v}_t}} m_{t-1} + \frac{(1-\beta_{1,t})}{\sqrt{\hat{v}_t}} g_t \right)$$

매개변수 벡터 $\vartheta \in \mathbb{R}^d$의 i^{th} 차원에 초점을 맞춥니다. 스칼라 $\vartheta_{*,i}$를 빼고 위의 업데이트 규칙의 양쪽을 제곱하면 다음과 같습니다,

$$(\vartheta_{t+1,i} - \vartheta_i^*)^2 = (\vartheta_{t,i} - \vartheta_i^*)^2 - \frac{2\alpha_t}{1-\beta_1^t} \left(\frac{\beta_{1,t}}{\sqrt{\hat{v}_{t,i}}} m_{t-1,i} + \frac{(1-\beta_{1,t})}{\sqrt{\hat{v}_{t,i}}} g_{t,i} \right)(\vartheta_{t,i} - \vartheta_i^*) + \alpha_t^2 \left(\frac{\hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}}} \right)^2$$

위의 방정식을 재정렬하고 영의 부등식 $ab \le a^2/2 + b^2/2$를 사용할 수 있습니다. 또한 다음과 같을 수 있습니다.

는 $\sqrt{\hat{v}_{t,i}} = \sqrt{\sum_{j=1}^{t} (1-\beta_2)\beta_2^{t-j} g_{j,i}^2} / \sqrt{1-\beta_2^t} \le \|g_{1:t,i}\|_2$ 및 $\beta_{1,t} \le \beta_1$ 그런 다음

$$g_{t,i}(\vartheta_{t,i} - \vartheta_i^*) = \frac{(1-\beta_1^t)\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_1)} \left[(\vartheta_{t,i} - \vartheta_i^*)^2 - (\vartheta_{t+1,i} - \vartheta_i^*)^2 \right] + \frac{\beta_{1,t}}{2\alpha_{t-1}(1-\beta_1)} (\vartheta_i^* - \vartheta_{t,i})^2 \sqrt{\hat{v}_{t-1,i}}$$
$$+ \frac{\beta_1 \alpha_{t-1}}{2(1-\beta_1)} \frac{m_{t-1,i}^2}{\sqrt{\hat{v}_{t-1,i}}} + \frac{\alpha_t}{2(1-\beta_1)} \frac{m_{t,i}^2}{\sqrt{\hat{v}_{t,i}}}$$

위의 부등식에 정리 10.4를 적용하여 $f_t(\vartheta_t) - f_t(\vartheta^*)$의 상한과 $t \in 1, \ldots, d$ 에대한 볼록함수의 수열에서 $i \in 1, \ldots, T$에 대한 모든 차원을 합산하여 후회 바운드를 유도한다:

$$R(T) \le \sum_{i=1}^{d} \frac{1}{2\alpha_1(1-\beta_1)} (\vartheta_{1,i} - \vartheta_i^*)^2 \sqrt{\hat{v}_{1,i}} + \sum_{i=1}^{d} \sum_{t=2}^{T} \frac{1}{2(1-\beta_1)} (\vartheta_{t,i} - \vartheta_i^*)^2 \left(\frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t} - \frac{\sqrt{\hat{v}_{t-1,i}}}{\alpha_{t-1}} \right)$$
$$+ \frac{\beta_1 \alpha G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^{d} \|g_{1:T,i}\|_2 + \frac{\alpha G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^{d} \|g_{1:T,i}\|_2$$

$$+\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\beta 1,t}{2\alpha_t(1-\beta)_{1,t}}(\vartheta_{,*}-\vartheta_{t,i})^2\sqrt{\hat{v}_{t,i}}$$

가정에서 $\|\vartheta_t - \vartheta^*\|_2 \le D$, $\|\theta_m - \vartheta_n\|_\infty \le D_\infty$ 를 구할 수 있습니다:

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

$$\le \frac{D_2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T v\hat{}T,i} + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^{d} \|g_{1:T,i}\|_2$$

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

마지막 항에 산술 기하 급수 상한을 사용할 수 있습니다:

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

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

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

따라서 다음과 같은 아쉬움이 남습니다:

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

□