Open Problem:
Black-Box Reductions & Adaptive Gradient Methods

Xinyi Chen Elad Hazan

Abstract

An open problem is described: reduce offline nonconvex stochastic optimization to regret minimization in online convex optimization. The conjectured reduction would explain the success of adaptive gradient methods for deep learning. A prize of 500$ is offered to the winner.

1 Motivation

Adaptive gradient methods are the most widely used optimization algorithms for training of deep neural networks [1]. The theory for these algorithms, starting from Adagrad [5, 10], is based on regret minimization in the context of online convex optimization [7].

The regret bounds of Adagrad can be better or worse than stochastic gradient descent, up to square root of the dimension factor, depending on the data. This can be very significant, as the dimension in deep neural network training is extremely large, and could potentially explain the performance improvements.

However, the regret guarantees of Adagrad imply faster convergence only for convex optimization. For nonconvex optimization, a different reduction from regret minimization to optimization is required, and this is the subject of this open problem.

2 The question

You are given an algorithm for online convex optimization $\mathcal{A}$, that has a guaranteed worst case regret bound [2]. Given a convex constraint set $\mathcal{K} \subseteq \mathbb{R}^d$ and an arbitrary sequence of convex cost functions $f_1, .., f_T : \mathcal{K} \mapsto \mathbb{R}$, the algorithm guarantees that

$$\sum_t f_t(\mathbf{x}_t) - \min_{\mathbf{x}^* \in \mathcal{K}} \sum_t f_t(\mathbf{x}^\star) \ \leq \ \text{Regret}_T(\mathcal{A}).$$

For example, the online gradient descent algorithm guarantees $\text{Regret}_T(OGD) \ \leq \ \frac{3}{2}GD\sqrt{T}$, where D is the diameter of $\mathcal{K}$, and G is an upper bound on the Lipschitz constant of the loss functions $\{f_t\}$.

The problem is to design a black box reduction from offline stochastic nonconvex optimization with a guaranteed performance that depends on the regret of $\mathcal{A}$. The performance metric we are looking for is average gradient norm across iterations, i.e. finding an approximate stationary point. More precisely, given a nonconvex β-smooth function $f : \mathcal{K} \mapsto \mathbb{R}$, we would like to obtain a sequence of points $\mathbf{x}_1, ..., \mathbf{x}_T \in \mathcal{K}$ such that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \ \leq \ O\left(\frac{\sqrt{\beta} \cdot \text{Regret}_T(A)}{T}\right). \tag{1}$$

Notice that this theorem implies the convergence rate of stochastic gradient descent for smooth nonconvex optimization, as per Theorem 3.4 in [6], when using the online gradient descent algorithm and its regret bound.

[1] at the time of writing, the Adam algorithm is one of the most widely cited research in the history of science [1]

[2] we assume this regret bound is deterministic for simplicity, but generalizations can be considered.

Replacing the regret bound of Adagrad in the conjectured equation (1) would give a similar rate in terms of number of iterations, i.e. $\frac{1}{\sqrt{T}}$. However, in terms of the dimension, the regret bound of Adagrad would imply convergence up to $\sqrt{d}$ faster! This would also apply to Adam [9] and basically most adaptive gradient methods whose theory is based on regret in online convex optimization.

3 Existing and recent progress

A reduction similar to the one proposed was put forth in previous works [11, 12, 2]. An algorithmic template from the reduction given in [2] is presented in Algorithm 1.

Algorithm 1 Non-convex to convex reduction

Input: OCO algorithm $\mathcal{A}$, β-smooth objective f, stochastic gradient oracle, parameters λ, w
for $t = 1$ to T **do**
  Let $f_t(\mathbf{x}) = f(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_t\|^2$
  Apply $\mathcal{A}$ to obtain $\mathbf{x}_{t+1}$ after w steps of the algorithm starting from x_t, using stochastic gradient oracle.
end for
return $\mathbf{x}_t$ such that $t = \arg\min_t \|\nabla f(\mathbf{x}_t)\|$.

Theorem A.2 in [2] [3] gives a provable convergence rate for this reduction in terms of the regret of the OCO algorithm $\mathcal{A}$. It is, however, unsatisfactory since the regret is taken over parts of the sequence rather than the entire sequence, and other subtleties.

Notable progress has been made recently in the work of [4], who give a black-box reduction with provable guarantees. However, their approach implies a bound similar to that of Equation (1) but with the adaptive regret notion [8] rather than regret.

A similar approach was attempted in [3], where the following procedure was studied:

Algorithm 2 NC2C reduction, second variant

Input: OCO algorithm $\mathcal{A}$, β-smooth objective function f, stochastic gradient oracle.
for $t = 1$ to T **do**
  Let f_t be the stochastic loss (e.g. i.i.d. sampled example).
  Let $\tilde{f}_t(\mathbf{x}) = f_t(\mathbf{x}) + \beta\|\mathbf{x} - \mathbf{x}_{t-1}\|^2$
  Apply single step of $\mathcal{A}$ to obtain $\mathbf{x}_{t+1} = \mathcal{A}(\tilde{f}_1, .., .\tilde{f}_t)$.
end for
return $\mathbf{x}_t$ such that $t = \arg\min_t \|\nabla f(\mathbf{x}_t)\|$.

A similar bound to (1) can be obtained, but this time with the dynamic regret notion rather than regret. The appeal of this reduction is that applying it with Adagrad (or Adam, or any other adaptive gradient method), is exactly running it on the nonconvex function directly with stochastic gradients.

4 NC2C2 Guarantee

Let $\text{DynamicRegret}_{\mathcal{A}}(f_{1:T}, \hat{x}_{1:T})$ denote the expected dynamic regret of algorithm $\mathcal{A}$ over a sequence of functions, $f_1, ..., f_t$, under the sequence of comparators $(\hat{x}_1, \ldots, \hat{x}_T)$, i.e.

$$\text{DynamicRegret}_{\mathcal{A}}(f_{1:T}, \hat{x}_{1:T}) = \sum_{t=1}^{T} f_t(x_t) - f_t(\hat{x}_t).$$

[3] which is sometimes referred to in the COLT community as "Naman's Lemma".

Lemma 1. *Suppose the function f satisfies $f(\mathbf{x}) - f(\mathbf{y}) \leq M \quad \forall\ \mathbf{x}, \mathbf{y}$, and let $\mathbf{x}_t^*$ denote a minimizer of $\bar{f}_t(\mathbf{x}) = f(\mathbf{x}) + \beta\|\mathbf{x} - \mathbf{x}_{t-1}\|^2$. Then the iterates $\mathbf{x}_t$ in Algorithm 2 satisfy*

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|^2\right] \leq \frac{6\beta}{T}\left(M + \mathit{DynamicRegret}_{\mathcal{A}}(\tilde{f}_{1:T}, \mathbf{x}^*_{1:T})\right),$$

where the expectation is taken over the randomness of the examples and the algorithm $\mathcal{A}$.

Proof. Let $\mathbb{E}_t$ denote conditioning on all randomness up to time t, then we have the following descent lemma,

$$\begin{aligned}
\mathbb{E}_t\left[f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t)\right] &= \mathbb{E}_t\left[\bar{f}_t(\mathbf{x}_{t-1}) - \bar{f}_t(\mathbf{x}_t) + \beta\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2^2\right] \\
&= \mathbb{E}_{t,f_t}\left[\tilde{f}_t(\mathbf{x}_{t-1}) - \tilde{f}_t(\mathbf{x}_t) + \beta\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2^2\right] \\
&\geq \mathbb{E}_{t,f_t}\left[\tilde{f}_t(\mathbf{x}_{t-1}) - \tilde{f}_t(\mathbf{x}_t^*) - (\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_t^*))\right] \\
&= \bar{f}(\mathbf{x}_{t-1}) - \bar{f}_t(\mathbf{x}_t^*) - \mathbb{E}_{\mathbf{x}_t,f_t}\left[(\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_t^*))\right] \\
&\geq \frac{1}{6\beta}\|\nabla f(\mathbf{x}_{t-1})\|^2 - \mathbb{E}_{\mathbf{x}_t,f_t}\left[\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_t^*)\right]. \qquad \text{(smoothness)}
\end{aligned}$$

Rearranging,

$$\|\nabla f(\mathbf{x}_{t-1})\|^2 \leq 6\beta\left(\mathbb{E}_t[f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t)] + \mathbb{E}_{t,f_t}\left[\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_t^*)\right]\right).$$

Summing up over the iterations and taking an unconditional expectation,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|^2\right] &\leq 6\beta\mathbb{E}\left[\sum_{t=2}^{T+1} f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t) + (\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_t^*))\right] \\
&\leq 6\beta\left(M + \mathbb{E}\left[\sum_{t=2}^{T+1}\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_t^*)\right]\right).
\end{aligned}$$

Thus the average gradient norm satisfies

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|^2\right] &\leq 6\beta\left(\frac{M}{T} + \frac{1}{T}\mathbb{E}\left[\sum_{t=2}^{T+1}\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_t^*)\right]\right) \\
&\leq 6\beta\left(\frac{M}{T} + \frac{1}{T}\mathrm{DynamicRegret}_{\mathcal{A}}(\tilde{f}_{1:T}, \mathbf{x}^*_{1:T})\right).
\end{aligned}$$

□

5 The Prize

We offer 500$ for an efficient black-box reduction giving the bound (1).

References

[1] Google scholar reveals its most influential papers for 2020. `https://www.nature.com/nature-index/news/google-scholar-reveals-most-influential-papers-research-citations-twenty-twenty`.

[2] Naman Agarwal, Brian Bullins, Xinyi Chen, Elad Hazan, Karan Singh, Cyril Zhang, and Yi Zhang. Efficient full-matrix adaptive regularization. In *International Conference on Machine Learning*, pages 102–110. PMLR, 2019.

[3] Xinyi Chen, Samory Kpotufe, and Elad Hazan. Convex to nonconvex reductions. In *unpublised manuscript*, 2023.

[4] Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. *arXiv preprint arXiv:2302.03775*, 2023.

[5] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[6] Elad Hazan. Lecture notes: Optimization for machine learning. *arXiv preprint arXiv:1909.03550*, 2019.

[7] Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[8] Elad Hazan and Comandur Seshadhri. Efficient learning algorithms for changing environments. In *Proceedings of the 26th annual international conference on machine learning*, pages 393–400, 2009.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

[11] Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaid Harchaoui. Catalyst for gradient-based nonconvex optimization. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 613–622. PMLR, 09–11 Apr 2018.

[12] Weiran Wang and Nathan Srebro. Stochastic nonconvex optimization with large minibatches. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 857–882. PMLR, 22–24 Mar 2019.