Proceedings of Machine Learning Research vol 196:1–8, 2024 37th Annual Conference on Learning Theory

Open Problem: Black-Box Reductions & Adaptive Gradient Methods for Nonconvex Optimization

Xinyi Chen XINYIC@PRINCETON.EDU
Elad Hazan EHAZAN@PRINCETON.EDU
Princeton University and Google DeepMind, Princeton, New Jersey

Editors: Shipra Agrawal and Aaron Roth

Abstract

We describe an open problem: reduce offline nonconvex stochastic optimization to regret minimization in online convex optimization. The conjectured reduction aims to make progress on explaining the success of adaptive gradient methods for deep learning. A prize of $500 is offered to the winner.

1. Introduction

Adaptive gradient methods are the most widely used optimization algorithms for training deep neural networks.[1] The theory for these algorithms, starting from AdaGrad (Duchi et al., 2011; McMahan and Streeter, 2010), is often rooted in regret minimization in the context of online convex optimization (OCO) (Hazan et al., 2016).

The regret bounds of AdaGrad in the convex setting can be better or worse than those of stochastic gradient descent (SGD), up to the square root of the dimension factor, depending on the data. This advantage can be very significant, as the dimension in deep neural network training is extremely large, and it could explain the performance improvements of adaptive optimizers in training deep neural networks.

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.

1.1. Related work

The analysis of adaptive gradient methods spans thousands of publications, and we omit a detailed survey in this open problem statement. Some recent advancements in the analysis of AdaGrad and related methods appear in Ward et al. (2019); Li and Orabona (2019); Zaheer et al. (2018); Défossez et al. (2022); Faw et al. (2022); Zhou et al. (2020). These results directly analyze the convergence of adaptive gradient methods to stationary points, rather than by reduction from regret in OCO.

There are a few benefits of a black-box reduction compared to a direct analysis:

- The convex world is simpler to analyze and we have tight regret bounds for many settings, in contrast to the more complex and general nonconvex optimization landscape.
- The best known regret bounds for OCO would imply faster rates for AdaGrad (and related methods) in certain scenarios than known by direct analysis.

1. At the time of writing, the Adam algorithm is one of the most widely cited research in the history of science (scholar article)

2. Problem statement

Assume we are given an algorithm $\mathcal{A}$ for OCO, 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

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x}^* \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}^\star) \ \leq \ \text{Regret}_T(\mathcal{A}).$$

For example, the online gradient descent (OGD) algorithm guarantees $\text{Regret}_T(OGD) \ \leq \ \frac{3}{2}GD\sqrt{T}$, where D is the diameter of $\mathcal{K}$ in ℓ_2 norm, and G is an upper bound on the Lipschitz constant of $\{f_t\}_{t=1}^{T}$. We are interested in the regret guarantee's dependence on the time horizon and dimension.

The problem is to design a black box-reduction from offline stochastic nonconvex optimization to OCO with a guaranteed performance that depends on the regret of $\mathcal{A}$. The performance metric we target is the average gradient norm across iterations, which is also the metric for finding an approximate stationary point. We make the following assumptions on the nonconvex objective function $f : \mathbb{R}^d \mapsto \mathbb{R}$, and for the sequel, $\|\cdot\|$ denotes the ℓ_2 norm of a vector.

Assumption 1 *The function f is β-smooth: it is differentiable and for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|.$$

Assumption 2 *We are given a starting point x_1, such that the distance in function value to optimality satisfies $f(x_1) - f(x^*) \leq M$.*

We have access to a stochastic gradient oracle $\mathcal{O}$ that is unbiased and has bounded variance.

Definition 3 *Let $\mathcal{O} : \mathbb{R}^d \to \mathbb{R}^d$ denote the stochastic gradient oracle, where given x, $\mathcal{O}$ outputs a gradient estimator $\tilde{\nabla} f(x)$: $\mathcal{O}(x) = \tilde{\nabla} f(x)$.*

Assumption 4 *For any $x \in \mathbb{R}^d$, $\mathbb{E}[\tilde{\nabla} f(x)] = \nabla f(x)$, and $\mathbb{E}[\|\tilde{\nabla} f(x) - \nabla f(x)\|^2] \leq \sigma^2$.*

We would like to obtain a sequence of points $\mathbf{x}_1, ..., \mathbf{x}_T \in \mathbb{R}^d$ such that

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

2.1. Why is this open problem important?

The convergence rate of SGD for smooth nonconvex optimization is known to be $O\left(\sigma\sqrt{\frac{\beta M}{T}}\right)$ (Ghadimi and Lan, 2013), where σ^2 is an upper bound on the variance, $\mathbb{E}[\|\tilde{\nabla} f(x) - \nabla f(x)\|^2]$. Note that this quantity implicitly depends on the problem dimension. The conjecture recovers this rate in the special case where the gradients are sampled from a stationary distribution and using the regret bound of the lazy variant of Online Gradient Descent (Hazan and Kale, 2010), as detailed in

2. We assume this regret bound is deterministic for simplicity, but generalizations can be considered similarly.

section A.1. Plugging in the regret bound of AdaGrad in the conjectured equation (1) would give a similar rate in terms of the number of iterations, i.e. $\frac{1}{\sqrt{T}}$, which is known to be tight (Arjevani et al., 2023). However, in terms of the dimension dependence, the regret bound of Adagrad would imply convergence up to $\sqrt{d}$ faster in certain cases due to a different measure of the diameter of the decision set. The variance term will instead scale as $D_\infty \mathrm{Tr}(\mathrm{diag}(\sum_{t=1}^T \nabla f(x_t) \nabla f(x_t)^\top)^{1/2})$, where D_∞ is the infinity-norm diameter of some decision set. This would also apply to Adam (Kingma and Ba, 2014) and most other adaptive gradient methods whose theory is based on regret in OCO.

We survey partial progress and reductions that are close in spirit to the required bound, but they do not obtain the desired provable speedup.

2.2. Constrained vs. unconstrained optimization

For nonconvex optimization over a constrained set, it is in general computationally hard to find a point with small gradient norm (Hazan et al., 2017). Therefore, the standard assumption in nonconvex optimization is unbounded domain and bounded function value difference between the initial point and the minimum. Alternatively, other solutions have been proposed such as bounding the projected gradient (Hazan et al., 2017).

3. Existing and recent progress

A reduction similar to the one proposed was put forth in previous works (Paquette et al., 2018; Wang and Srebro, 2019; Agarwal et al., 2019). This reduction is episodic, and an algorithmic template from Agarwal et al. (2019) is presented in Algorithm 1.

Algorithm 1 Nonconvex to convex reduction

Input: OCO algorithm $\mathcal{A}$, β-smooth objective f, stochastic gradient oracle $\mathcal{O}$, parameters λ, w
for $k = 1$ to K **do**
  Let $f_k(\mathbf{x}) = f(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{x}_k\|^2$ be the regularized loss of the k-th epoch.
  Apply $\mathcal{A}$ to obtain $\mathbf{x}_{k+1}$ after w steps of the algorithm starting from $\mathbf{x}_k$, using $\mathcal{O}$.
end for
return $\mathbf{x}_{k^*}$ such that $k^* = \mathrm{argmin}_k \|\nabla f(\mathbf{x}_k)\|$.

Theorem A.2 of Agarwal et al. (2019) [3] gives a provable convergence rate for this reduction in terms of the number of stochastic oracle calls (Kw) and the regret[4] 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 by Cutkosky et al. (2023), who give a black-box reduction with provable guarantees. Instead of regularizing the losses to be strongly convex, their reduction leverages online learning over linearized losses to ”predict” an update direction. However, their approach implies a bound close to that of Equation (1) but with the adaptive regret notion of Hazan and Seshadhri (2009) rather than regret.

3. It is sometimes referred to in the COLT community as “Naman’s Lemma”.
4. The actual lemma is phrased in terms of optimization performance, but it can be rephrased in terms of regret.

A non-episodic variant of Algorithm 1 was attempted by Chen et al. (2023), where the following procedure was studied in the finite-sum setting:

Algorithm 2 Nonconvex to convex reduction, second variant

Input: OCO algorithm $\mathcal{A}$, β-smooth objective function $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$.
for $t = 1$ to T **do**
  Let $f_t(\mathbf{x})$ be the loss function defined by the t-th batch of examples.
  Let $\tilde{f}_t(\mathbf{x}) = f_t(\mathbf{x}) + \beta\|\mathbf{x} - \mathbf{x}_{t-1}\|^2$ be the strongly convex regularized loss.
  Apply a 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 $t^* = \text{argmin}_t \|\nabla f(\mathbf{x}_t)\|$.

A similar bound to (1) can be obtained, but with the dynamic regret notion instead regret. Dynamic regret is intimately related to adaptive regret, as adaptive regret can be reduced to dynamic regret. This reduction is perhaps the most direct, in the sense that applying it with AdaGrad (or Adam or any other adaptive gradient method) is exactly running it on the regularized nonconvex function with stochastic gradients. The guarantee is stated in the lemma below, and the proof is in the appendix.

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

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

We further assume that we have independent batches at every time step, a standard assumption in stochastic optimization.

Assumption 5 *The set of batches that we receive at each time step is dependent of each other, and also of the iterates x_t.*

Lemma 6 *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*

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

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

It is more subtle to bound the dynamic regret, and it is more difficult to derive optimization guarantees using this notion.

4. The Prize

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

Acknowledgments

We thank Samory Kpotufe, Andras Gyorgy, and Dale Schuurmans for helpful discussions.

References

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.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

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

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.

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

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

Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 313–355. PMLR, 02–05 Jul 2022.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi: 10.1137/120880811.

Elad Hazan and Satyen Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80:165–188, 2010.

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.

Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pages 1433–1441. PMLR, 2017.

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

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

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 983–992. PMLR, 16–18 Apr 2019.

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

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.

scholar article. Google scholar reveals its most influential papers for 2020.

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.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6677–6686. PMLR, 09–15 Jun 2019.

Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization, 2020.

Appendix A. Proof of Lemma 6

Proof Let $\mathbb{E}_{\tilde{f}_t}$ denote the expectation taken over the random batch in the definition of $\tilde{f}_t$. We have the following descent lemma:

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

Rearranging,

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

Summing up over the iterations and taking an unconditional expectation,

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|^2\right] &\leq 6\beta\mathbb{E}\left[\sum_{t=2}^{T+1}\left[f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t) + (\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_t^*))\right]\right] \\ &\leq 6\beta\left(f(\mathbf{x}_1) - f(\mathbf{x}_{T+1}) + \mathbb{E}\left[\sum_{t=2}^{T+1}\tilde{f}_t(\mathbf{x}_t) - \tilde{f}_t(\mathbf{x}_t^*)\right]\right). \end{aligned}$$

Thus the average gradient norm satisfies

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

■

A.1. Applying the reduction with OGD

Recall that the open problem requests the following bound:

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

Suppose we have linear losses with loss vectors $g_1, \ldots, g_T$, the lazy FTRL algorithm with Euclidean regularization has the following regret bound from Theorem 3 of Hazan and Kale (2010),

$$\text{Regret}_T(\text{Lazy-OGD}) = D\min_{\mu\in\mathbb{R}^d}\sqrt{\sum_{t=1}^{T}\|g_t - \mu\|^2}.$$

Notice that if $g_1, \ldots, g_T$ are the linearized losses, and if they come from a stationary distribution, eg. we are at a stationary point, then

$$\mathbb{E}[\text{Regret}_T(\text{Lazy-OGD})] \leq D\sigma\sqrt{T}.$$

We obtain:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \;\leq\; O\left(\sigma\sqrt{\frac{M\beta}{T}}\right).$$