Escaping Saddle Points with Adaptive Gradient Methods

Matthew Staib [1 2] **Sashank Reddi** [3] **Satyen Kale** [3] **Sanjiv Kumar** [3] **Suvrit Sra** [1]

Abstract

Adaptive methods such as Adam and RMSProp are widely used in deep learning but are not well understood. In this paper, we seek a crisp, clean and precise characterization of their behavior in nonconvex settings. To this end, we first provide a novel view of adaptive methods as preconditioned SGD, where the preconditioner is estimated in an online manner. By studying the preconditioner on its own, we elucidate its purpose: it rescales the stochastic gradient noise to be isotropic near stationary points, which helps escape saddle points. Furthermore, we show that adaptive methods can efficiently estimate the aforementioned preconditioner. By gluing together these two components, we provide the first (to our knowledge) second-order convergence result for any adaptive method. The key insight from our analysis is that, compared to SGD, adaptive methods escape saddle points faster, and can converge faster overall to second-order stationary points.

1. Introduction

Stochastic first-order methods are the algorithms of choice for training deep networks, or more generally optimization problems of the form $\operatorname{argmin}_x \mathbb{E}_z[f(x,z)]$. While vanilla stochastic gradient descent (SGD) is still the most popular such algorithm, there has been much recent interest in adaptive methods that adaptively change learning rates for each parameter. This is a very old idea, e.g. (Jacobs, 1988); modern variants such as Adagrad (Duchi et al., 2011; McMahan & Streeter, 2010) Adam (Kingma & Ba, 2014) and RMSProp (Tieleman & Hinton, 2012) are widely used in deep learning due to their good empirical performance.

Adagrad uses the square root of the sum of the outer product of the past gradients to achieve adaptivity. At time step t, Adagrad updates the parameters in the following manner:

$$x_{t+1} = x_t - G_t^{-1/2} g_t,$$

where g_t is a noisy stochastic gradient at x_t and $G_t = \sum_{i=1}^t g_i g_i^T$. More often, a diagonal version of Adagrad is used due to practical considerations, which effectively yields a per parameter learning rate. In the convex setting, Adagrad achieves provably good performance, especially when the gradients are sparse. Although Adagrad works well in sparse convex settings, its performance appears to deteriorate in (dense) nonconvex settings. This performance degradation is often attributed to the rapid decay of the learning rate in Adagrad over time, which is a consequence of rapid increase in eigenvalues of the matrix G_t.

To tackle this issue, variants of Adagrad such as Adam and RMSProp have been proposed, which replace the sum of the outer products with an exponential moving average i.e., $G_t = (1-\beta)\sum_{i=1}^t \beta^{t-i} g_i g_i^T$ for some constant $\beta \in (0,1)$. This connection with Adagrad is often used to justify the design of Adam and RMSProp (e.g. (Goodfellow et al., 2016)). Although this connection is simple and appealing, it is clearly superficial. For instance, while learning rates in Adagrad decrease monotonically, it is not necessarily the case with Adam or RMSProp as shown recently in Reddi et al. (2018b), leading to their non-convergence in even simple convex settings. Thus, a principled understanding of these adaptive methods is largely missing.

In this paper, we introduce a much simpler way of thinking about adaptive methods such as Adam and RMSProp. Roughly, adaptive methods try to precondition SGD by some matrix A, e.g. when A is diagonal, A_{ii} corresponds to the effective stepsize for coordinate i. For some choices of A the algorithms do not have oracle access to A, but instead form an estimate $\hat{A} \approx A$. We separate out these two steps, by 1) giving convergence guarantees for an idealized setting where we have access to A, then 2) proving bounds on the quality of the estimate $\hat{A}$. Our approach makes it possible to effectively intuit about the algorithms, prove convergence guarantees (including second-order convergence), and give insights about how to choose algorithm parameters. It also leads to a number of surprising results, including an understanding of why the Reddi et al. (2018b) counterexample is hard for adaptive methods, why adaptive methods tend to escape saddle points faster than SGD (observed empirically

[1]MIT EECS [2]Based on work performed at Google Research New York [3]Google Research New York. Correspondence to: Matthew Staib <mstaib@mit.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

in (Reddi et al., 2018a)), insights into how to tune Adam's parameters, and (to our knowledge) the first *second-order* convergence proof for any adaptive method.

Contributions: In addition to the aforementioned novel viewpoint, we also make the following key contributions:

- We develop a new approach for analyzing convergence of adaptive methods leveraging the preconditioner viewpoint and by way of disentangling estimation from the behavior of the *idealized* preconditioner.
- We provide *second-order convergence* results for adaptive methods, and as a byproduct, first-order convergence results. To the best of our knowledge, ours is the first work to show second order convergence for any adaptive method.
- We provide theoretical insights on how adaptive methods escape saddle points quickly. In particular, we show that the preconditioner used in adaptive methods leads to isotropic noise near stationary points, which helps escape saddle points faster.
- Our analysis also provides practical suggestions for tuning the exponential moving average parameter β.

1.1. Related work

There is an immense amount of work studying nonconvex optimization for machine learning, which is too much to discuss here in detail. Thus, we only briefly discuss two lines of work that are most relevant to our paper here. First, the recent work e.g. (Chen et al., 2018; Reddi et al., 2018b; Zou et al., 2018) to understand and give theoretical guarantees for adaptive methods such as Adam and RMSProp. Second, the technical developments in using first-order algorithms to achieve nonconvex second-order convergence (see Definition 2.1) e.g. (Ge et al., 2015; Allen-Zhu & Li, 2018; Jin et al., 2017; Lee et al., 2016).

Nonconvex convergence of adaptive methods. Many recent works have investigated convergence properties of adaptive methods. However, to our knowledge, all these results either require convexity or show only first-order convergence to stationary points. Reddi et al. (2018b) showed non-convergence of Adam and RMSProp in simple convex settings and provided a variant of Adam, called AMSGrad, with guaranteed convergence in the convex setting; Zhou et al. (2018) generalized this to a nonconvex first-order convergence result. Zaheer et al. (2018) showed first-order convergence of Adam when the batch size grows over time. Chen et al. (2018) bound the nonconvex convergence rate for a large family of Adam-like algorithms, but they essentially need to assume the effective stepsize is well-behaved (as in AMSGrad). Agarwal et al. (2018) give a convex convergence result for a full-matrix version of RMSProp, which they extend to the nonconvex case via iteratively optimizing convex functions. Their algorithm uses a fixed sliding window instead of an exponential moving average. Mukkamala & Hein (2017) prove improved convergence bounds for Adagrad in the online strongly convex case; they prove similar results for RMSProp, but only in a regime where it is essentially the same as Adagrad. Ward et al. (2018) give a nonconvex convergence result for a variant of Adagrad which employs an adaptively decreasing single learning rate (not per-parameter). Zou et al. (2018) give sufficient conditions for first-order convergence of Adam.

Nonconvex second order convergence of first order methods. Starting with Ge et al. (2015) there has been a resurgence in interest in giving first-order algorithms that find *second* order stationary points of nonconvex objectives, where the gradient is small and the Hessian is nearly positive semidefinite. Most other results in this space operate in the deterministic setting where we have exact gradients, with carefully injected isotropic noise to escape saddle points. Levy (2016) show improved results for normalized gradient descent. Some algorithms rely on Hessian-vector products instead of pure gradient information e.g. (Agarwal et al., 2017; Carmon et al., 2018); it is possible to reduce Hessian-vector based algorithms to gradient algorithms (Xu et al., 2018; Allen-Zhu & Li, 2018). Jin et al. (2017) improve the dependence on dimension to polylogarithmic. Mokhtari et al. (2018) work towards adapting these techniques for constrained optimization. Most relevant to our work is that of Daneshmand et al. (2018), who prove convergence of SGD with better rates than Ge et al. (2015). Concurrent with our paper, Fang et al. (2019) give even better rates for SGD. Our work differs in that we provide second-order results for *preconditioned* SGD.

2. Notation and definitions

The objective function is f, and the gradient and Hessian of f are ∇f and $H = \nabla^2 f$, respectively. Denote by $x_t \in \mathbb{R}^d$ the iterate at time t, by g_t an unbiased stochastic gradient at x_t and by ∇_t the expected gradient at t. The matrix G_t refers to $\mathbb{E}[g_t g_t^T]$. Denote by $\lambda_{\max}(G)$ and $\lambda_{\min}(G)$ the largest and smallest eigenvalues of G, and $\kappa(G)$ is the condition number $\lambda_{\max}(G)/\lambda_{\min}(G)$ of G. For a vector v, its elementwise p-th power is written v^p. The objective $f(x)$ has global minimizer x^*, and we write $f^* = f(x^*)$. The Euclidean norm of a vector v is written as $\|v\|$, while for a matrix M, $\|M\|$ refers to the operator norm of M. The matrix I is the identity matrix, whose dimension should be clear from context.

Definition 2.1 (Second-order stationary point)**.** A (τ_g, τ_h)-stationary point of f is a point x so that $\|\nabla f(x)\| \leq \tau_g$ and

Algorithm 1 Preconditioned SGD

Input: initial x_0, time T, stepsize η, preconditioner $A(x)$
for $t = 0, \dots, T$ **do**
  $g_t \leftarrow$ stochastic gradient at x_t
  $A_t \leftarrow A(x_t)$ ▷ e.g. $A_t = \mathbb{E}[g_t g_t^T]^{-1/2}$
  $x_{t+1} \leftarrow x_t - \eta A_t g_t$
end for

Algorithm 2 Full-matrix RMSProp

Input: initial x_0, time T, stepsize η, small number $\varepsilon > 0$ for stability
for $t = 0, \dots, T$ **do**
  $g_t \leftarrow$ stochastic gradient
  $\hat{G}_t = \beta \hat{G}_{t-1} + (1-\beta) g_t g_t^T$
  $A_t = (\hat{G}_t + \varepsilon I)^{-1/2}$
  $x_{t+1} \leftarrow x_t - \eta A_t g_t$
end for

$\lambda_{\min}(\nabla^2 f(x)) \geq -\tau_h$, where $\tau_g, \tau_h > 0$.

As is standard (e.g. Nesterov & Polyak (2006)), we will discuss only $(\tau, \sqrt{\rho\tau})$-stationary points, where ρ is the Lipschitz constant of the Hessian.

3. The RMSProp Preconditioner

Recall that methods like Adam and RMSProp replace the running sum $\sum_{i=1}^t g_i g_i^T$ used in Adagrad with an exponential moving average (EMA) of the form $(1-\beta)\sum_{i=1}^t \beta^{t-i} g_i g_i^T$, e.g. full-matrix RMSProp is described formally in Algorithm 2. One key observation is that $\hat{G}_t = (1-\beta)\sum_{i=1}^t \beta^{t-i} g_i g_i^T \approx \mathbb{E}[g_t g_t^T] =: G_t$ if β is chosen appropriately; in other words, at time t, the accumulated $\hat{G}_t$ can be seen as an approximation of the true second moment matrix $G_t = \mathbb{E}[g_t g_t^T]$ at the current iterate. Thus, RMSProp can be viewed as preconditioned SGD (Algorithm 1) with the preconditioner being $A_t = G_t^{-1/2}$. In practice, it is too expensive to compute G_t exactly since it requires summing over all training samples. Practical adaptive methods (see Algorithm 2) estimate this preconditioner (or a diagonal approximation) on-the-fly via an EMA.

Before developing our formal results, we will build intuition about the behavior of adaptive methods by studying an idealized adaptive method (IAM) with perfect access to G_t. In the rest of this section, we make use of idealized RMSProp to answer some simple questions about adaptive methods that we feel have not yet been addressed satisfactorily.

3.1. What is the purpose of the preconditioner?

Why should preconditioning by $A = \mathbb{E}[gg^T]^{-1/2}$ help optimization? The original Adam paper (Kingma & Ba, 2014) argues that Adam is an approximation to natural gradient descent, since if the objective f is a log-likelihood, $\mathbb{E}[gg^T]$ approximates the Fisher information matrix, which captures curvature information in the space of distributions. There are multiple issues with comparing adaptive methods to natural gradient descent, which we discuss in Appendix A. Instead, Balles & Hennig (2018) argue that the primary function of adaptive methods is to equalize the stochastic gradient noise in each direction. But it is still *not* clear why or how equalized noise should help optimization.

Our IAM abstraction makes it easy to explain precisely how rescaling the gradient noise helps. Specifically, we manipulate the update rule for idealized RMSProp:

$$x_{t+1} \leftarrow x_t - \eta A_t g_t \tag{1}$$
$$= x_t - \eta A_t \nabla_t - \eta \underbrace{A_t(g_t - \nabla_t)}_{=:\xi_t} \tag{2}$$

The $A_t\nabla_t$ term is deterministic; only ξ_t is stochastic, with mean $\mathbb{E}[A_t(g_t - \nabla_t)] = A_t\,\mathbb{E}[g_t - \nabla_t] = 0$. Take $\varepsilon = 0$ and assume $G_t = \mathbb{E}[g_t g_t^T]$ is invertible, so that $\xi_t = G_t^{-1/2}(g_t - \nabla_t)$. Now we can be more precise about how RMSProp rescales gradient noise. Specifically, we compute the covariance of the noise ξ_t:

$$\mathrm{Cov}(\xi_t) = I - G_t^{-1/2}\nabla_t \nabla_t^T G_t^{-1/2}. \tag{3}$$

The key insight is: near stationary points, ∇_t will be small, so that the noise covariance $\mathrm{Cov}(\xi_t)$ is approximately the identity matrix I. In other words, at stationary points, the gradient noise is approximately isotropic. This observation hints at why adaptive methods are so successful for non-convex problems, where one of the main challenges is to escape saddle points (Reddi et al., 2018a). Essentially all first-order approaches for escaping saddlepoints rely on adding carefully tuned isotropic noise, so that regardless of what the escape direction is, there is enough noise in that direction to escape with high probability.

3.2. (Reddi et al., 2018b) counterexample resolution

Recently, Reddi et al. (2018b) provided a simple *convex* stochastic counterexample on which RMSProp and Adam do not converge. Their reasoning is that RMSProp and Adam too quickly forget about large gradients from the past, in favor of small (but poor) gradients at the present. In contrast, for RMSProp with the idealized preconditioner (Algorithm 1 with $A = \mathbb{E}[gg^T]^{-1/2}$), there is no issue, but the preconditioner A cannot be computed in practice. Rather, for this example, the exponential moving average estimation scheme fails to adequately estimate the preconditioner.

The counterexample is an optimization problem of the form

$$\min_{x \in [-1,1]} \quad F(x) = p f_1(x) + (1-p) f_2(x), \tag{4}$$

where the stochastic gradient oracle returns ∇f_1 with probability p and ∇f_2 otherwise. Let $\zeta > 0$ be "small," and $C >$

0 be "large." Reddi et al. (2018b) set $p = (1+\zeta)/(C+1)$, $f_1(x) = Cx$, and $f_2(x) = -x$. Overall, then, $F(x) = \zeta x$ which is minimized at $x = -1$, however Reddi et al. (2018b) show that RMSProp has $\mathbb{E}[F(x_t)] \geq 0$ and so incurs suboptimality gap at least ζ. In contrast, the idealized preconditioner is a function of

$$\mathbb{E}[g^2] = p\left(\frac{\partial f_1}{\partial x}\right)^2 + (1-p)\left(\frac{\partial f_2}{\partial x}\right)^2 = C(1+\zeta) - \zeta$$

which is a constant independent of x. Hence the preconditioner is constant, and, up to the choice of stepsize, idealized RMSProp on this problem is the same as SGD, which of course will converge.

The difficulty for practical adaptive methods (which estimate $\mathbb{E}[g^2]$ via an EMA) is that as C grows, the variance of the estimate of $\mathbb{E}[g^2]$ grows too. Thus Reddi et al. (2018b) break Adam by making estimation of $\mathbb{E}[g^2]$ harder.

4. Main Results: Gluing Estimation and Optimization

The key enabling insight of this paper is to separately study the preconditioner and its estimation via EMA, then combine these to give proofs for practical adaptive methods. We will prove a formal guarantee that the EMA estimate $\hat{G}_t$ is close to the true G_t. By combining our estimation results with the underlying behavior of the preconditioner, we will be able to give convergence proofs for practical adaptive methods that are constructed in a novel, modular way.

Separating these two components enables more general results: we actually analyze preconditioned SGD (Algorithm 1) with oracle access to an arbitrary preconditioner $A(x)$. Idealized RMSProp is but one particular instance. Our convergence results depend only on specific properties of the preconditioner $A(x)$, with which we can recover convergence results for many RMSProp variants simply by bounding the appropriate constants. For example, $A = (\mathbb{E}[gg^T]^{1/2} + \varepsilon I)^{-1}$ corresponds to full-matrix Adam with $\beta_1 = 0$ or RMSProp as commonly implemented. For cleaner presentation, we instead focus on the variant $A = (\mathbb{E}[gg^T] + \varepsilon I)^{-1/2}$, but our proof technique can handle either case or its diagonal approximation.

4.1. Estimating from Moving Sequences

The above discussion about IAM is helpful for intuition, and as a base algorithm for analyzing convergence. But it remains to understand how well the estimation procedure works, both for intuition's sake and for later use in a convergence proof. In this section we introduce an abstraction we name "estimation from moving sequences." This abstraction will allow us to guarantee high quality estimates of the preconditioner, or, for that matter, any similarly constructed preconditioner. Our results will moreover make apparent how to choose the β parameter in the exponential moving average: β should increase with the stepsize η. Increasing β over time has been supported both empirically (Shazeer & Stern, 2018) as well as theoretically (Mukkamala & Hein, 2017; Zou et al., 2018; Reddi et al., 2018b), though to our knowledge, the precise pinning of β to the stepsize η is new.

Suppose there is a sequence of states $x_1, x_2, \ldots, x_T \in \mathcal{X}$, e.g. the parameters of our model at each time step. We have access to the states x_t, but more importantly we know the states are not changing too fast: $\|x_t - x_{t-1}\|$ is bounded for all t. There is a Lipschitz function $G : \mathcal{X} \to \mathbb{R}^{d\times d}$, which in our case is the second moment matrix of the stochastic gradients, but could be more general. We would like to estimate $G(x)$ for each $x = x_t$, but we have only a noisy oracle $Y(x)$ for $G(x)$, which we assume is unbiased and has bounded variance. Our goal is, given noisy reads $Y_1, \ldots, Y_T$ of $G(x_1), \ldots, G(x_T)$, to estimate $G(x_T)$ at the current point x_T as well as possible.

We consider estimators of the form $\sum_{t=1}^T w_t Y_t$. For example. setting $w_T = 1$ and all others to zero would yield an unbiased (but high variance) estimate of $G(x_T)$. We could assign more mass to older samples Y_t, but this will introduce bias into the estimate. By optimizing this bias-variance tradeoff, we can get a good estimator. In particular, taking w to be an exponential moving average (EMA) of $\{Y_t\}_{t=1}^T$ will prioritize more recent and relevant estimates, while placing enough weight on old estimates to reduce the variance. The tradeoff is controlled by the EMA parameter β; e.g. if the sequence x_t moves slowly (the stepsize is small), we will want large β because older iterates are still very relevant.

In adaptive methods, the underlying function $G(x)$ we want to estimate is $\mathbb{E}[gg^T]$, and every stochastic gradient g gives us an unbiased estimate $Y = gg^T$. The diagonal approximation also fits this formulation: we can set $Y = \operatorname{diag}(g^2)$ as an unbiased estimate of the matrix $\mathbb{E}[\operatorname{diag}(g^2)]$. With these applications in mind, we formalize our results in terms of matrix estimation. By combining standard matrix concentration inequalities (e.g. Tropp (2015)) with bounds on how fast the sequence moves, we obtain the following result:

Theorem 4.1. *Assume $\|x_t - x_{t+1}\| \leq \eta M$. The function $G : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is matrix-valued and L-Lipschitz. For each t, Y_t is a random matrix with $\mathbb{E}[Y_t] = G(x_t)$, $\|G_t - \mathbb{E}[G_t]\| \leq R$, and $\|\mathbb{E}[(G_t - \mathbb{E}[G_t])^2]\| \leq \sigma_{\max}^2$. Set $w_t \propto \beta^{T-t}$ with $\sum_{t=1}^T w_t = 1$ and assume $T > 4/(1-\beta)$. Then with probability $1-\delta$, the estimation error $\Phi = \left\|\sum_{t=1}^T w_t Y_t - G(x_T)\right\|$ is bounded by*

$$\Phi \leq O(\sigma_{\max}\sqrt{1-\beta}\sqrt{\log(d/\delta)} + ML\eta/(1-\beta)).$$

This is optimized by $\beta = 1 - C\eta^{2/3}$, for which the bound is $O((\eta M \sigma_{\max}^2 (\log(d/\delta))L)^{1/3})$ as long as $T > C'\eta^{-2/3}$.

The proof is given in Appendix G. As long as T is sufficiently large, we can get a high quality estimate of $G_t = \mathbb{E}[g_t g_t^T]$. For this, it suffices to start off the underlying optimization algorithm with $W = O(\eta^{-2/3})$ burn-in iterations where our estimate is updated but the algorithm is not started. This burn-in period will not affect asymptotic runtime as long as $W = O(\eta^{-2/3}) = O(T)$. In our non-convex convergence results we will require $T = O(\tau^{-4})$ and $\eta = O(\tau^2)$, so that $W = O(\tau^{-4/3})$ which is much smaller than T. In practice, one can get away with much shorter (or no) burn-in period.

If β is properly tuned, while running an adaptive method like RMSProp, we will get good estimates of $G = \mathbb{E}[gg^T]$ from samples gg^T. However, we actually require a good estimate of $A = \mathbb{E}[gg^T]^{-1/2}$ and variants. To treat estimation in a unified way, we introduce estimable matrix sequences:

Definition 4.1. A $(W, T, \eta, \Delta, \delta)$-*estimable matrix sequence* is a sequence of matrices $\{A(x_t)\}_{t=1}^{W+T}$ generated from $\{x_t\}_t$ with $\|x_t - x_{t-1}\| \leq \eta$ so that with probability $1-\delta$, after a burn-in of time W, we can achieve an estimate sequence $\{\hat{A}_t\}$ so that $\|\hat{A}_t - A_t\| \leq \Delta$ simultaneously for all times $t = W+1, \ldots, W+T$.

Applying Theorem 4.1 and union bounding over all time $t = W+1, \ldots, W+T$, we may state a concise result in terms of Definition 4.1:

Proposition 4.1. *Suppose $G = \mathbb{E}[g_t g_t^T]$ is L-Lipschitz as a function of x. When applied to a generator sequence $\{x_t\}$ with $\|x_t - x_{t-1}\| \leq \eta M$ and samples $Y_t = g_t g_t^T$, the matrix sequence $G_t = \mathbb{E}[g_t g_t^T]$ is $(W, T, \eta M, \Delta, \delta)$-estimable with $W = O(\eta^{-2/3})$, $T = \Omega(W)$, and $\Delta = O(\eta^{1/3}\sigma_{\max}^{2/3}(\log(2Td/\delta)^{1/3}M^{1/3}L^{1/3})$.*

We are hence guaranteed a good estimate of G. What we actually want, though, is a good estimate of the preconditioner $A = (G + \varepsilon I)^{-1/2}$. In Appendix H we show how to bound the quality of an estimate of A. One simple result is:

Proposition 4.2. *Suppose $G = \mathbb{E}[gg^T]$ is L-Lipschitz as a function of x. Further suppose a uniform bound $\lambda_{\min}(G)I \preceq G(x)$ for all x, with $\lambda_{\min}(G) > 0$. When applied to a generator sequence $\{x_t\}$ with $\|x_t - x_{t-1}\| \leq \eta M$ and samples $Y_t = g_t g_t^T$, the matrix sequence $A_t = (G_t + \varepsilon I)^{-1/2}$ is $(W, T, \eta M, \Delta, \delta)$-estimable with $W = O(\eta^{-2/3})$, $T = \Omega(W)$, and $\Delta = O((\eta\sigma_{\max}^2 \log(2Td/\delta)ML)^{1/3}(\varepsilon + \lambda_{\min}(G))^{-3/2})$.*

4.2. Convergence Results

We saw in the last two sections that it is simple to reason about adaptive methods via IAM, and that it is possible to compute a good estimate of the preconditioner. But we still need to glue the two together in order to get a convergence proof for practical adaptive methods.

In this section we will give non-convex convergence results, first for IAM and then for practical realizations thereof. We start with first-order convergence as a warm-up, and then move on to second-order convergence. In each case we give a bound for IAM, study it, and then give the corresponding bound for practical adaptive methods.

4.2.1. Assumptions and notation

We want results for a wide variety of preconditioners A, e.g. $A = I$, the RMSProp preconditioner $A = (G + \varepsilon I)^{-1/2}$, and the diagonal version thereof, $A = (\mathrm{diag}(G) + \varepsilon I)^{-1/2}$. To facilitate this and the future extension of our approach to other preconditioners, we give guarantees that hold for general preconditioners A. Our bounds depend on A via the following properties:

Definition 4.2. We say $A(x)$ is a $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$-*preconditioner* if, for all x, the following bounds hold. First, $\|A\nabla f\|^2 \leq \Lambda_1 \|A^{1/2}\nabla f\|^2$. Second, if $\tilde{f}(x)$ is the quadratic approximation of f at some point x_0, we assume $\|A(\nabla f - \nabla\tilde{f})\| \leq \Lambda_2 \|\nabla f - \nabla \tilde{f}\|$. Third, $\Gamma \geq \mathbb{E}[\|Ag\|^2]$. Fourth, $\nu \leq \lambda_{\min}(A\,\mathbb{E}[gg^T]A^T)$. Finally, $\lambda_- \leq \lambda_{\min}(A)$.

Note that we could bound $\Lambda_1 = \Lambda_2 = \lambda_{\max}(A)$. but in practice Λ_1 and Λ_2 may be smaller, since they depend on the behavior of A only in specific directions. In particular, if the preconditioner A is well-aligned with the Hessian, as may be the case if the natural gradient approximation is valid, then Λ_1 would be very small. If f is exactly quadratic, Λ_2 can be taken as a constant. The constant Γ controls the magnitude of (rescaled) gradient noise, which affects stability at a local minimum. Finally, ν gives a lower bound on the amount of gradient noise in any direction; when ν is larger it is easier to escape saddle points[1]. For shorthand, a $(\cdot, \cdot, \Gamma, \cdot, \lambda_-)$-preconditioner needs to satisfy only the corresponding inequalities.

In Appendix C we provide bounds on these constants for variants of the second moment preconditioner. We highlight the two most relevant cases, for SGD and RMSProp:

Proposition 4.3. *The preconditioner $A = I$ is a $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$-preconditioner, with $\Lambda_1 = \Lambda_2 = 1$, $\Gamma \leq \mathbb{E}[\|g\|^2] \leq d \cdot \mathrm{tr}(G)$, $\nu \leq \lambda_{\min}(G)$, and $\lambda_- = 1$.*

Proposition 4.4. *The preconditioner $A = (G + \varepsilon I)^{-1/2}$ is a $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$-preconditioner, with*

$$\Lambda_1 = \Lambda_2 = \frac{1}{(\lambda_{\min}(G) + \varepsilon)^{1/2}}, \quad \Gamma = \frac{d\lambda_{\max}(G)}{\varepsilon + \lambda_{\max}(G)},$$

$$\nu = \frac{\lambda_{\min}(G)}{\lambda_{\min}(G) + \varepsilon}, \quad \text{and} \quad \lambda_- = (\lambda_{\max}(G) + \varepsilon)^{-1/2}.$$

[1] In cases where $G = \mathbb{E}[gg^T]$ is rank deficient, e.g. in high-dimensional finite sum problems, lower bounds on $\lambda_{\min}(G)$ should be understood as lower bounds on $\mathbb{E}[(v^T g)^2]$ for escape directions v from saddle points, analogous to the "CNC condition" from (Daneshmand et al., 2018).

4.2.2. First-order convergence

Proofs are given in Appendix F. For all first-order results, we assume that A is a $(\cdot, \cdot, \Gamma, \cdot, \lambda_-)$-preconditioner. The proof technique is essentially standard, with minor changes in order to accomodate general preconditioners. First, suppose we have exact oracle access to the preconditioner:

Theorem 4.2. *Run preconditioned SGD with preconditioner A and stepsize $\eta = \tau^2\lambda_-/(L\Gamma)$. For small enough τ, after $T = 2(f(x_0) - f^*)L\Gamma/(\tau^4\lambda_-^2)$ iterations,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \tau^2. \tag{5}$$

Now we consider an alternate version where instead of the preconditioner A_t, we precondition by an noisy version $\hat{A}_t$ that is close to A_t, i.e. $\|\hat{A}_t - A_t\| \leq \Delta$.

Theorem 4.3. *Suppose we have access to an inexact preconditioner $\hat{A}$, which satisfies $\|\hat{A} - A\| \leq \Delta$ for $\Delta < \lambda_-/2$. Run preconditioned SGD with preconditioner $\hat{A}$ and stepsize $\eta = \tau^2\lambda_-/(4\sqrt{2}L\Gamma)$. For small enough τ, after $T = 32(f(x_0) - f^*)L\Gamma/(\tau^4\lambda_-^2)$ iterations, we will have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \tau^2. \tag{6}$$

The results are the same up to constants. In other words, as long as we can achieve less than $\lambda_-/2$ error, we will converge at essentially the same rate as if we had the exact preconditioner. In light of this, for the second-order convergence results, we treat only the noisy version.

Theorem 4.3 gives a convergence bound assuming a good estimate of the preconditioner, and our estimation results guarantee a good estimate. By gluing together Theorem 4.3 with our estimation results for the RMSProp preconditioner, i.e. Proposition 4.2, we can give a convergence result for bona fide RMSProp:

Corollary 4.1. *Consider RMSProp with burn-in, as in Algorithm 3, where we estimate $A = (G + \varepsilon I)^{-1/2}$. Retain the same choice of $\eta = O(\tau^2)$ and $T = O(\tau^{-4})$ as in Theorem 4.3. For small enough τ, such a choice of η will yield $\Delta < \lambda_-/2$. Choose all other parameters e.g. β in accordance with Proposition 4.2. In particular, choose $W = \Theta(\eta^{-2/3}) = \Theta(\tau^{-4/3}) = O(T)$ for the burn-in parameter. Then with probability $1 - \delta$, in overall time $O(W + T) = O(\tau^{-4})$, we achieve*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \tau^2. \tag{7}$$

Algorithm 3 RMSProp with burn-in

Input: initial x_0, time T, stepsize η, burn-in length W
$\hat{G}_0 \leftarrow \text{BurnIn}(W, \beta)$ ▷ Appendix B
for $t = 0, \ldots, T$ **do**
  $g_t \leftarrow$ stochastic gradient
  $\hat{G}_t \leftarrow \beta\hat{G}_{t-1} + (1-\beta)g_t g_t^T$
  $\hat{A}_t \leftarrow \hat{G}_t^{-1/2}$
  $x_{t+1} \leftarrow x_t - \eta\hat{A}_t g_t$
end for

4.2.3. Second-order convergence

Now we leverage the power of our high level approach to prove nonconvex second-order convergence for adaptive methods. Like the first-order results, we start by proving convergence bounds for a generic, possibly inexact preconditioner A. Our proof is based on that of Daneshmand et al. (2018) for SGD, and therefore we achieve the same $O(\tau^{-5})$ rate. It may be possible to improve our result using the technique of Fang et al. (2019), which is concurrent work to ours. However, our focus is on the preconditioner, and our study of it is wholly new. Accordingly, we study the convergence of Algorithm 4, which is the same as Algorithm 1 (generic preconditioned SGD) except that once in a while we take a large stepsize so we may escape saddlepoints. The proof is given completely in Appendix E. At a high level, we show the algorithm makes progress when the gradient is large and when we are at a saddle point, and does not escape from local minima. Our analysis uses all the constants specified in Definition 4.2, e.g. the speed of escape from saddle points depends on ν, the lower bound on stochastic gradient noise.

Then, as before, we simply fuse our convergence guarantees with our estimation guarantees. The end result is, to our knowledge, the first nonconvex second-order convergence result for any adaptive method.

Definitions for second-order results. Assume further that the Hessian H is ρ-Lipschitz and the preconditioner $A(x)$ is α-Lipschitz. The dependence on these constants is made precise in the proof, in Appendix E. The usual stepsize is η, while r is the occasional large stepsize that happens every t_{thresh} iterations. We tolerate a small probability of failure δ. For all results, we assume A is a $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$-preconditioner. For simplicity, we assume the noisy estimate $\hat{A}$ also satisfies the Λ_1 inequality. We will also assume a uniform bound on $\|Ag\| \leq M = O(\sqrt{\Gamma})$.

The proofs rely on a few other quantities that we optimally determine as a function of the problem parameters: f_{thresh} is a threshold on the function value progress, and $g_{\text{thresh}} = f_{\text{thresh}}/t_{\text{thresh}}$ is the time-amortized average of f_{thresh}. We specify the precise values of all quantities in the proof.

Theorem 4.4. *Consider Algorithm 4 with inexact preconditioner $\hat{A}_t$ and exact preconditioner A_t satisfying the*

Algorithm 4 Preconditioned SGD with increasing stepsize

Input: initial x_0, time T, stepsizes η, r, threshold t_{thresh}, matrix error Δ
for $t = 0, \ldots, T$ **do**
  $A_t \leftarrow A(x_t)$ ▷ preconditioner at x_t
  $\hat{A}_t \leftarrow$ any matrix with $\|\hat{A}_t - A_t\| \leq \Delta$
  $g_t \leftarrow$ stochastic gradient at x_t
  if t mod $t_{\text{thresh}} = 0$ **then**
    $x_{t+1} \leftarrow x_t - r\hat{A}_t g_t$
  else
    $x_{t+1} \leftarrow x_t - \eta\hat{A}_t g_t$
  end if
end for

preceding requirements. Suppose that for all t, we have $\|\hat{A}_t - A_t\| = O(\tau^{1/2})$*. Then for small* τ*, with probability* $1-\delta$*, we reach an* $(\tau, \sqrt{\rho\tau})$*-stationary point in time*

$$T = \tilde{O}\left(\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} \cdot \frac{L^3}{\rho\delta^3} \cdot \tau^{-5}\right). \tag{8}$$

The big-O suppresses other constants given in the proof.

To prove a result for bona fide RMSProp, we need to combine Theorem 4.4 with an algorithm that maintains a good estimate of $G = \mathbb{E}[gg^T]$ (and consequently $A = (G + \varepsilon I)^{-1/2}$). This is more delicate than the first-order case because now the stepsize varies. Whenever we take a large stepsize, the estimation algorithm will need to hallucinate S number of smaller steps in order to keep the estimate accurate. Our overall scheme is formalized in Appendix B, for which the following convergence result holds:

Corollary 4.2. *Consider the RMSProp version of Algorithm 4 that is described in Appendix B. Retain the same choice of* $\eta = O(\tau^{5/2})$*,* $r = O(\tau)$*, and* $T = O(\tau^{-5})$ *as in Theorem 4.4. For small enough* τ*, such a choice of* η *will yield* $\Delta < \lambda_-/2$*. Choose* $W = \Theta(\eta^{-2/3}) = \Theta(\tau^{-5/3}) = O(T)$ *for the burn-in parameter Choose* $S = O(\tau^{-3/2})$*, so that as far as the estimation scheme is concerned, the stepsize is bounded by* $\max\{\eta, r/S\} = O(\tau^{5/2}) = O(\eta)$*. Then as before, with probability* $1-\delta$*, we can reach an* $(\tau, \sqrt{\rho\tau})$*-stationary point in total time*

$$W + T = \tilde{O}\left(\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} \cdot \frac{L^3}{\rho\delta^3} \cdot \tau^{-5}\right), \tag{9}$$

where $\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-$ *are the constants describing* $A = (G + \varepsilon I)^{-1/2}$.

Again, as in the first order results, one could substitute in any other estimable preconditioner. In particular, in Appendix D we discuss the more common diagonal version of RMSProp.

5. Discussion

Separating the estimation step from the preconditioning enables evaluation of different choices for the preconditioner.

5.1. How to set the regularization parameter ε

In the adaptive methods literature, it is still a mystery how to properly set the regularization parameter ε that ensures invertibility of $G + \varepsilon I$. When the optimality tolerance τ is small enough, estimating the preconditioner is not the bottleneck. Thus, focusing only on the idealized case, one could just choose ε to minimize the bound. Our first-order results depend on ε only through the following term:

$$\frac{\Gamma}{\lambda_{\min}(A)} \leq \frac{d\lambda_{\min}(G)}{\varepsilon + \lambda_{\min}(G)} \cdot (\lambda_{\max}(G) + \varepsilon), \tag{10}$$

where we have used the preconditioner bounds from Proposition 4.4. This is minimized by taking $\varepsilon \to \infty$, which suggests using identity preconditioner, or SGD. In contrast, for second-order convergence, the bound is

$$\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} \leq d^4\kappa(G)^4(\lambda_{\max}(G) + \varepsilon), \tag{11}$$

which is instead minimized with $\varepsilon = 0$. So for the best second-order convergence rate, it is desireable to set ε as small as possible. Note that since our bounds hold only for small enough convergence tolerance τ, it is possible that the optimal ε should depend in some way on τ.

5.2. Comparison to SGD

Another important question we make progress towards is: when are adaptive methods better than SGD? Our second-order result depends on the preconditioner only through $\Lambda_1^4\Lambda_2^4\Gamma^4/(\lambda_-^{10}\nu^4)$. Plugging in Proposition 4.3 for SGD, we may bound

$$\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} \leq \frac{\mathbb{E}[\|g\|^2]^4}{\lambda_{\min}(G)^4} \leq d^4\kappa(G)^4, \tag{12}$$

while for full-matrix RMSProp, we have

$$\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} \leq d^4\kappa(G)^4(\lambda_{\max}(G) + \varepsilon). \tag{13}$$

Setting $\varepsilon = 0$ for simplicity, we conclude that full-matrix RMSProp converges faster if $\lambda_{\max}(G) \leq 1$.

Now suppose that for a given optimization problem, the preconditioner A is well-aligned with the Hessian so that $\Lambda_1 = O(1)$ (e.g. if the natural gradient approximation holds) and that near saddle points the objective is essentially quadratic so that $\Lambda_2 = O(1)$. In this regime, the preconditioner dependence of idealized full matrix RMSProp is $d^4\lambda_{\max}(G)^5$, which yields a better result than SGD when $\lambda_{\max}(G) \leq \lambda_{\min}(G)^{-4}$. This will happen whenever $\lambda_{\min}(G)$ is relatively small. Thus, when there is not much noise in the escape direction, and the Hessian and $G^{-1/2}$ are not poorly aligned, RMSProp will converge faster overall. Similar phenomenon can be shown for the diagonal case when the approximation is good, per the results in Appendix C and D.

5.3. Alternative preconditioners

Our analysis inspires the design of other preconditioners: e.g., if at each iteration we sample two independent stochastic gradients g_1 and g_2, we have unbiased sample access to $(g_1 - g_2)(g_1 - g_2)^T$, which in expectation yields the covariance $\Sigma = \mathrm{Cov}(g)$ instead of the second moment matrix of g. It immediately follows that we can prove second-order convergence results for an algorithm that constructs an exponential moving average estimate of Σ and preconditions by $\Sigma^{-1/2}$, as advocated by Ida et al. (2017).

5.4. Tuning the EMA parameter β

Another mystery of adaptive methods is how to set the exponential moving average (EMA) parameter β. In practice β is typically set to a constant, e.g. 0.99, while other parameters such as the stepsize η are tuned more carefully and may vary over time. While our estimation guarantee Theorem 4.1, suggests setting $\beta = 1 - O(\eta^{2/3})$, the specific formula depends on constants that may be unknown, e.g. Lipschitz constants and gradient norms. Instead, one could set $\beta = 1 - C\eta^{2/3}$, and search for a good choice of the hyperparameter C. For example, the common initial choice of $\eta = 0.001$ and $\beta = 0.99$ corresponds to $C = 1$.

6. Experiments

We experimentally test our claims about adaptive methods escaping saddle points, and our suggestion for setting β.

Escaping saddle points. First, we test our claim that when the gradient noise is ill-conditioned, adaptive methods escape saddle points faster than SGD, and often converge faster to (approximate) local minima. We construct a two dimensional[2] non-convex problem $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ where $f_i(x) = \frac{1}{2}x^T H x + b_i^T x + \|x\|_{10}^{10}$. Here, $H = \mathrm{diag}([1, -0.1])$, so f has a saddle point at the origin with objective value zero. The vectors b_i are chosen so that sampling b uniformly from $\{b_i\}_{i=1}^{n}$ yields $\mathbb{E}[b] = 0$ and $\mathrm{Cov}(b) = \mathrm{diag}([1, 0.01])$. Hence at the origin there is an escape direction but little gradient noise in that direction.

We initialize SGD and (diagonal) RMSProp (with $\beta = 1 - \eta^{2/3}$) at the saddle point and test several stepsizes η for each. Results for the first 10^4 iterations are shown in Figure 1. In order to escape the saddle point as fast as RMSProp, SGD requires a substantially larger stepsize, e.g. SGD needs $\eta = 0.01$ to escape as fast as RMSProp does with $\eta = 0.001$. But with such a large stepsize, SGD cannot converge to a small neighborhood of the local minimum, and instead bounces around due to gradient noise. Since RMSProp can escape with a small stepsize, it can converge to a much smaller neighborhood of the local minimum. Overall, for any fixed final convergence criterion, RMSProp escapes faster and converges faster overall.

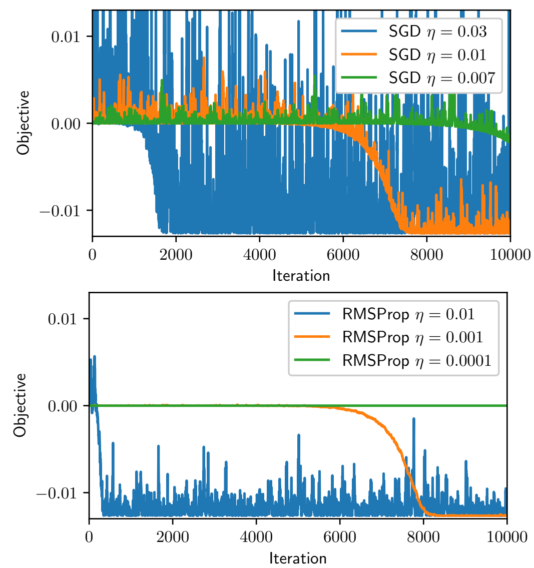

Figure 1. SGD (top) vs RMSProp (bottom) performance escaping a saddle point with poorly conditioned gradient noise. Compared to RMSProp, SGD requires a larger stepsize to escape as quickly, which negatively impacts convergence to the local minimum.

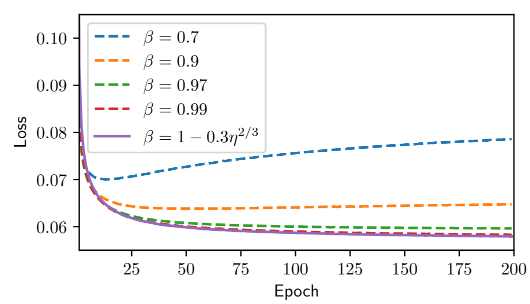

Figure 2. Performance on MNIST logistic regression of RMSProp with different choices of β and decreasing stepsize.

Setting the EMA parameter β. Next, we test our recommendations regarding setting the EMA parameter β. We consider logistic regression on MNIST. We use (diagonal) RMSProp with batch size 100, decreasing stepsize $\eta_t = 0.001/\sqrt{t}$ and $\varepsilon = 10^{-8}$, and compare different schedules for β. Specifically we test $\beta \in \{0.7, 0.9, 0.97, 0.99\}$ (so that $1 - \beta$ is spaced roughly logarithmically) as well as our recommendation of $\beta_t = 1 - C\eta_t^{2/3}$ for $C \in \{0.1, 0.3, 1\}$. As shown in Figure 2, all options for β have similar performance initially, but as η_t decreases, large β yields substantially better performance. In particular, our decreasing β schedule achieved the best performance, and moreover was insensitive to how C was set.

[2] The same phenomenon still holds in higher dimensions but the presentation is simpler with $d = 2$.

Acknowledgements

SS acknowledges support from the DARPA Lagrange grant, an Amazon Research Award, and the NSF-CAREER Award (ID 1846088). We thank Nicolas Le Roux for helpful conversations.

References

Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2017, pp. 1195–1199, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4528-6. doi: 10.1145/3055399.3055464.

Agarwal, N., Bullins, B., Chen, X., Hazan, E., Singh, K., Zhang, C., and Zhang, Y. The case for full-matrix adaptive regularization. *arXiv preprint arXiv:1806.02958*, 2018.

Allen-Zhu, Z. and Li, Y. Neon2: Finding local minima via first-order oracles. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 3720–3730. Curran Associates, Inc., 2018.

Balles, L. and Hennig, P. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Carmon, Y., Duchi, J., Hinder, O., and Sidford, A. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018. doi: 10.1137/17M1114296.

Chen, X., Liu, S., Sun, R., and Hong, M. On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

Daneshmand, H., Kohler, J., Lucchi, A., and Hofmann, T. Escaping saddles with stochastic gradients. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1155–1164, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011. ISSN 1532-4435.

Fang, C., Lin, Z., and Zhang, T. Sharp analysis for nonconvex sgd escaping from saddle points. *arXiv preprint arXiv:1902.00247*, 2019.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points — online stochastic gradient for tensor decomposition. In Grnwald, P., Hazan, E., and Kale, S. (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 797–842, Paris, France, 03–06 Jul 2015. PMLR.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Ida, Y., Fujiwara, Y., and Iwamura, S. Adaptive learning rate via covariance matrix based preconditioning for deep neural networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pp. 1923–1929, 2017. doi: 10.24963/ijcai.2017/267.

Jacobs, R. A. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1724–1732, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

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

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In Feldman, V., Rakhlin, A., and Shamir, O. (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1246–1257, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

Levy, K. Y. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

McMahan, H. B. and Streeter, M. J. Adaptive bound optimization for online convex optimization. In *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pp. 244–256, 2010.

Mokhtari, A., Ozdaglar, A., and Jadbabaie, A. Escaping saddle points in constrained optimization. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 3633–3643. Curran Associates, Inc., 2018.

Mukkamala, M. C. and Hein, M. Variants of RMSProp and Adagrad with logarithmic regret bounds. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2545–2553, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Nesterov. *Introductory lectures on convex optimization : a basic course*. Kluwer Academic Publishers, Boston, 2004. ISBN 1402075537.

Nesterov, Y. and Polyak, B. T. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Reddi, S., Zaheer, M., Sra, S., Poczos, B., Bach, F., Salakhutdinov, R., and Smola, A. A generic approach for escaping saddle points. In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1233–1242, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018a. PMLR.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018b.

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4596–4604, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Tieleman, T. and Hinton, G. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4 (2):26–31, 2012.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015. ISSN 1935-8237. doi: 10.1561/2200000048.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *arXiv preprint arXiv:1806.01811*, 2018.

Xu, Y., Rong, J., and Yang, T. First-order stochastic algorithms for escaping from saddle points in almost linear time. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 5531–5541. Curran Associates, Inc., 2018.

Zaheer, M., Reddi, S., Sachan, D., Kale, S., and Kumar, S. Adaptive methods for nonconvex optimization. In *NIPS*. 2018.

Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W. A sufficient condition for convergences of Adam and RMSProp. *arXiv preprint arXiv:1811.09358*, 2018.

A. More Insights from Idealized Adaptive Methods (IAM)

Suppose for now that we have oracle access to $G_t = \mathbb{E}[g_t g_t^T]$. Why should preconditioning by $A = \mathbb{E}[gg^T]^{-1/2}$ help optimization? The original Adam paper (Kingma & Ba, 2014) argues that Adam is an approximation to natural gradient descent, since if the objective f is a log-likelihood, $\mathbb{E}[gg^T]$ approximates the Fisher information matrix F, which captures curvature information in the space of distributions. This connection is tenuous at best, since the approximation $F \approx \mathbb{E}[gg^T]$ is only valid near optimality. Moreover, the exponent is wrong: Adam preconditions by $\mathbb{E}[gg^T]^{-1/2}$, but natural gradient should precondition by $\mathbb{E}[gg^T]^{-1}$. But using the exponent -1 is reported in the literature as unstable, even for Adagrad: "*without the square root operation, the algorithm performs much worse*" (Ruder, 2016). So the exponent is changed to $-1/2$ instead of -1.

Both of the above issues with the natural gradient interpretation are also pointed out in Balles & Hennig (2018), who argue that the primary function of adaptive methods is to equalize the stochastic gradient noise in each direction. But it is still *not* clear precisely why or how equalized noise should help optimization.

By assuming oracle access to $\mathbb{E}[gg^T]$, we can immediately argue that the exponent cannot be more aggressive than $-1/2$. Suppose we run preconditioned SGD with the preconditioner G_t^{-1} (instead of $G_t^{-1/2}$ as in RMSProp), and apply this to a noiseless problem; that is, g_t always equals the full gradient $\nabla_t = \nabla f(x_t)$. The preconditioner is then

$$A_t = (\mathbb{E}[g_t g_t^T] + \varepsilon I)^{-1} = (\nabla_t \nabla_t^T + \varepsilon I)^{-1}. \tag{14}$$

Taking $\varepsilon \to 0$, the idealized RMSProp update approaches

$$x_{t+1} \leftarrow x_t - \eta \frac{\nabla_t}{\|\nabla_t\|^2}. \tag{15}$$

First, the actual descent direction is not changed, and curvature is totally absent. Second, the resulting algorithm is unstable unless η decreases rapidly: as x_t approaches a stationary point, the magnitude of the step $\nabla_t/\|\nabla_t\|^2$ grows arbitrarily large, making it impossible to converge without rapidly decreasing the stepsize.

By contrast, using the standard $-1/2$ exponent and taking $\varepsilon \to 0$ in the noiseless case yields normalized gradient descent:

$$x_{t+1} \leftarrow x_t - \eta \frac{\nabla_t}{\|\nabla_t\|}. \tag{16}$$

In neither case do adaptive methods actually change the direction of descent (e.g. via curvature information); only the stepsize is changed.

B. Algorithm Details

Per our estimation results in Section 4.1, we must alter RMSProp to ensure it achieves an accurate estimate of the preconditioner. Namely, before updating the parameter x_t, we need to burn-in the estimate for several iterations so the initial estimate $\hat{G}_0$ is accurate. This subroutine is given in Algorithm 5.

Later, when we prove second-order convergence, we need to modify RMSProp to occassionally take a large step. However, this complicates estimation: per Theorem 4.1, estimation quality deteriorates as the step size increases. Naively applying Theorem 4.1 to the large stepsize yields an estimate of G that is not accurate enough. To get around this, every time RMSProp takes a large step, we will hallucinate a number of smaller steps to feed into the estimation procedure. This is formalized in Algorithm 6. Overall, the variant of RMSProp we study is formalized in Algorithm 7.

C. Curvature and noise constants for different preconditioners

Our analysis for general preconditioners depends on constants $\Lambda_1, \Lambda_2, \Gamma, \nu$, as well as $\lambda_- = \lambda_{\min}(A)$ that measure various properties of the preconditioner A. For convenience, we reproduce the definition:

Definition C.1. We say $A(x)$ is a $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$-*preconditioner* if, for all x in the domain, the following bounds hold. First, $\|A\nabla f\|^2 \leq \Lambda_1 \|A^{1/2}\nabla f\|^2$. Second, if $\tilde{f}(x)$ is the quadratic approximation of f at some point x_0, we assume $\|A(\nabla f - \nabla \tilde{f})\| \leq \Lambda_2 \|\nabla f - \nabla \tilde{f}\|$. Third, $\Gamma \geq \mathbb{E}[\|Ag\|^2]$. Fourth, $\nu \leq \lambda_{\min}(A\,\mathbb{E}[gg^T]A^T)$. Finally, $\lambda_- \leq \lambda_{\min}(A)$.

As before, we write $G = \mathbb{E}[gg^T]$ throughout.

Algorithm 5 BurnIn

function BURNIN(burn-in length W, β)
  for $t = 0, \dots, W$ **do**
    $g_t \leftarrow$ stochastic gradient
    $\hat{G}_t \leftarrow \beta \hat{G}_{t-1} + (1-\beta) g_t g_t^T$
  end for
  return $\hat{G}_t$
end function

Algorithm 6 Hallucinate

function HALLUCINATE(hallucination length S, β, $\hat{G}$, x_{start}, x_{end})
  for $s = 0, \dots, S$ **do**
    $g_s \leftarrow$ stochastic gradient at $x_{\text{start}} + \frac{s}{S}(x_{\text{end}} - x_{\text{start}})$
    $\hat{G} \leftarrow \beta \hat{G} + (1-\beta) g_s g_s^T$
  end for
  return $\hat{G}$
end function

C.1. Constants for identity preconditioner

In the simplest case, $A = I$ and we merely run SGD. We reproduce Proposition 4.3:

Proposition C.1. *The preconditioner* $A = I$ *is a* $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$*-preconditioner, with* $\Lambda_1 = \Lambda_2 = 1$, $\Gamma = \mathbb{E}[\|g\|^2]$, $\nu \leq \lambda_{\min}(G)$, *and* $\lambda_- = 1$.

The overall second-order complexity depends on

$$\frac{\Lambda_1 \Lambda_2 \Gamma}{\nu} = \frac{\mathbb{E}[\|g\|^2]}{\lambda_{\min}(G)}, \tag{17}$$

as well as $\lambda_- = \lambda_{\min}(A) = 1$.

Proof of Proposition C.1. Clearly, $\Lambda_1 = \Lambda_2 = \lambda_- = 1$. Then,

$$\mathbb{E}[\|Ag\|^2] = \mathbb{E}[\|g\|^2] =: \Gamma. \tag{18}$$

Finally,

$$\lambda_{\min}(AGA^T) = \lambda_{\min}(G) =: \nu. \tag{19}$$

□

Algorithm 7 Full-matrix RMSProp with increasing stepsize

Input: initial x_0, time T, stepsizes η, r, threshold t_{thresh}, time S, burn-in length W, momentum β
$\hat{G}_0 \leftarrow$ BURNIN(W, β) ▷ Algorithm 5
for $t = 0, \dots, T$ **do**
  $g_t \leftarrow$ stochastic gradient at x_t
  $\hat{G}_t \leftarrow \beta \hat{G}_{t-1} + (1-\beta) g_t g_t^T$
  $A_t \leftarrow \hat{G}_t^{-1/2}$
  if $t \bmod t_{\text{thresh}} = 0$ **then**
    $x_{t+1} \leftarrow x_t - r A_t g_t$
    $\hat{G}_t \leftarrow$ HALLUCINATE$(S, \beta, \hat{G}_t, x_t, x_{t+1})$ ▷ Algorithm 6
  else
    $x_{t+1} \leftarrow x_t - \eta A_t g_t$
  end if
end for

C.2. Constants for full matrix IAM

Write $G = \mathbb{E}[gg^T]$, and define the preconditioner A by $A = (G + \varepsilon I)^{-1/2}$. We reproduce Proposition 4.4:

Proposition C.2. *The preconditioner $A = (G + \varepsilon I)^{-1/2}$ is a $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$-preconditioner, with*

$$\Lambda_1 = \Lambda_2 = (\lambda_{\min}(G) + \varepsilon)^{-1/2},\ \Gamma = \frac{d\lambda_{\max}(G)}{\varepsilon + \lambda_{\max}(G)},\ \nu = \frac{\lambda_{\min}(G)}{\lambda_{\min}(G) + \varepsilon}, \tag{20}$$

and $\lambda_- = (\lambda_{\max}(G) + \varepsilon)^{-1/2}$.

Overall, the complexity depends on $\Lambda_1\Lambda_2\Gamma/\nu$:

$$\frac{\Lambda_1\Lambda_2\Gamma}{\nu} = \frac{1}{\sqrt{\lambda_{\min}(G) + \varepsilon}} \cdot \frac{1}{\sqrt{\lambda_{\min}(G) + \varepsilon}} \cdot \frac{d\lambda_{\max}(G)}{\varepsilon + \lambda_{\max}(G)} \cdot \frac{\lambda_{\min}(G) + \varepsilon}{\lambda_{\min}(G)} \tag{21}$$

$$= \frac{d\lambda_{\max}(G)}{(\varepsilon + \lambda_{\max}(G))\lambda_{\min}(G)}. \tag{22}$$

Therefore

$$\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} \leq \left(\frac{d\lambda_{\max}(G)}{(\varepsilon + \lambda_{\max}(G))\lambda_{\min}(G)}\right)^4 (\lambda_{\max}(G) + \varepsilon)^5 \tag{23}$$

$$= d^4\kappa(G)^4(\lambda_{\max}(G) + \varepsilon) \tag{24}$$

Note that when $\varepsilon = 0$ and we do not regularize the preconditioner, the complexity bound is

$$\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} = d^4\kappa(G)^4\lambda_{\max}(G). \tag{25}$$

If we make the optimistic but often reasonable assumptions that $\Lambda_1 = O(1)$ (if A is aligned well with the Hessian) and $\Lambda_2 = O(1)$ (the function f is essentially quadratic at saddle points) then all dependence on $\lambda_{\min}(G)$ vanishes, and the bound is

$$\frac{\Gamma^4}{\lambda_-^{10}\nu^4} = d^4\lambda_{\max}(G)^5. \tag{26}$$

Proof of Proposition 4.4. We can bound both Λ_1 and Λ_2 by

$$\Lambda_1, \Lambda_2 \leq \lambda_{\max}(A) = \lambda_{\min}(G + \varepsilon I)^{-1/2} = (\lambda_{\min}(G) + \varepsilon)^{-1/2}. \tag{27}$$

For Γ, we need to bound $\mathbb{E}[\|Ag\|^2] = \operatorname{tr}(A^2G)$. Expanding, we may write

$$A^2G = (G + \varepsilon I)^{-1}G. \tag{28}$$

The mapping $t \mapsto t/(t + \varepsilon)$ is increasing, so by using the bound $\lambda_{\max}(G)I \succeq G$, we may bound

$$A^2G \preceq \frac{\lambda_{\max}(G)}{\varepsilon + \lambda_{\max}(G)}I. \tag{29}$$

It follows that we can bound the trace of A^2G by

$$\Gamma = d \cdot \frac{\lambda_{\max}(G)}{\varepsilon + \lambda_{\max}(G)}. \tag{30}$$

Next, ν is a bound on the least eigenvalue of

$$AGA^T = (G + \varepsilon I)^{-1/2}G(G + \varepsilon I)^{-1/2} = (G + \varepsilon I)^{-1}G. \tag{31}$$

Since $t \mapsto t/(t + \varepsilon)$ is increasing, it is minimized when t is small. Therefore

$$\lambda_{\min}(AGA^T) \geq \frac{\lambda_{\min}(G)}{\lambda_{\min}(G) + \varepsilon} =: \nu. \tag{32}$$

□

C.3. Constants for diagonal IAM

Define the preconditioner A by $A = \operatorname{diag}(\mathbb{E}[g^2] + \varepsilon)^{-1/2}$.

Proposition C.3. *The preconditioner* $A = \operatorname{diag}(\mathbb{E}[g^2] + \varepsilon)^{-1/2}$ *is a* $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$*-preconditioner, with*

$$\Lambda_1 = \Lambda_2 = \left(\varepsilon + \min_{i\in[d]} \mathbb{E}[g_i^2]\right)^{-1/2}, \; \Gamma = \frac{d \max_{i\in[d]} \mathbb{E}[g_i^2]}{\varepsilon + \max_{i\in[d]} \mathbb{E}[g_i^2]}, \; \nu = \frac{\lambda_{\min}(G \operatorname{diag}(G)^{-1}) \cdot \min_{i\in[d]} \mathbb{E}[g_i^2]}{\varepsilon + \min_{i\in[d]} \mathbb{E}[g_i^2]}, \tag{33}$$

and $\lambda_- = (\varepsilon + \max_{i\in[d]} \mathbb{E}[g_i]^2)^{-1/2}$.

Overall,

$$\frac{\Lambda_1\Lambda_2\Gamma}{\nu} = \frac{\varepsilon + \min_{i\in[d]} \mathbb{E}[g_i^2]}{(\varepsilon + \min_{i\in[d]} \mathbb{E}[g_i^2]) \cdot \lambda_{\min}(G \operatorname{diag}(G)^{-1}) \min_{i\in[d]} \mathbb{E}[g_i^2]} \cdot \frac{d \cdot \max_{i\in[d]} \mathbb{E}[g_i^2]}{\varepsilon + \max_{i\in[d]} \mathbb{E}[g_i^2]} \tag{34}$$

$$= \frac{1}{\lambda_{\min}(G \operatorname{diag}(G)^{-1}) \min_{i\in[d]} \mathbb{E}[g_i^2]} \cdot \frac{d \cdot \max_{i\in[d]} \mathbb{E}[g_i^2]}{\varepsilon + \max_{i\in[d]} \mathbb{E}[g_i^2]} \tag{35}$$

$$\tag{36}$$

so the overall second-order dependence is

$$\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} = \frac{(\varepsilon + \max_{i\in[d]} \mathbb{E}[g_i^2])^5}{\lambda_{\min}(G \operatorname{diag}(G)^{-1})^4 (\min_{i\in[d]} \mathbb{E}[g_i^2])^4} \cdot \frac{d^4 \cdot (\max_{i\in[d]} \mathbb{E}[g_i^2])^4}{(\varepsilon + \max_{i\in[d])} \mathbb{E}[g_i^2])^4} \tag{37}$$

$$= \frac{(\varepsilon + \max_{i\in[d]} \mathbb{E}[g_i^2]) \cdot d^4 \cdot (\max_{i\in[d]} \mathbb{E}[g_i^2])^4}{\lambda_{\min}(G \operatorname{diag}(G)^{-1})^4 (\min_{i\in[d]} \mathbb{E}[g_i^2])^4}. \tag{38}$$

If we set $\varepsilon = 0$ and do not regularize the preconditioner, the complexity bound is

$$\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} = \frac{d^4 \cdot (\max_{i\in[d]} \mathbb{E}[g_i^2])^5}{\lambda_{\min}(G \operatorname{diag}(G)^{-1})^4 (\min_{i\in[d]} \mathbb{E}[g_i^2])^4}. \tag{39}$$

Proof of Proposition C.3. As before, we can bound both Λ_1 and Λ_2 by

$$\Lambda_1, \Lambda_2 \leq \lambda_{\max}(A) = \left(\varepsilon + \min_{i\in[d]} \mathbb{E}[g_i^2]\right)^{-1/2}. \tag{40}$$

For Γ, using the same manipulations as before, we want to bound

$$\mathbb{E}[\|Ag\|^2] = \operatorname{tr}(\operatorname{diag}(\mathbb{E}[g^2]) \operatorname{diag}(\varepsilon + \mathbb{E}[g^2])^{-1}) \tag{41}$$

$$= \operatorname{tr}\left(\operatorname{diag}\left(\frac{\mathbb{E}[g^2]}{\varepsilon + \mathbb{E}[g^2]}\right)\right) \tag{42}$$

$$\leq d \cdot \frac{\max_{i\in[d]} \mathbb{E}[g_i^2]}{\varepsilon + \max_{i\in[d]} \mathbb{E}[g_i^2]}. \tag{43}$$

Again, bounding ν is difficult, as we would need to bound the least eigenvalue of

$$A\,\mathbb{E}[gg^T]A = \mathbb{E}[gg^T] \operatorname{diag}(\varepsilon + \mathbb{E}[g^2])^{-1} \tag{44}$$

$$= G(\varepsilon + \operatorname{diag}(G))^{-1} \tag{45}$$

$$= G(\operatorname{diag}(G)^{-1} - \operatorname{diag}(G)^{-1}(\varepsilon^{-1}I + \operatorname{diag}(G)^{-1})^{-1} \operatorname{diag}(G)^{-1}) \tag{46}$$

$$= G \operatorname{diag}(G)^{-1}(I - (\varepsilon^{-1}I + \operatorname{diag}(G)^{-1})^{-1} \operatorname{diag}(G)^{-1}). \tag{47}$$

The first two terms are ν if we had not added ε to A. The remaining terms can be bounded as before by

$$I - (\varepsilon^{-1}I + \operatorname{diag}(G)^{-1})^{-1} \operatorname{diag}(G)^{-1} \succeq \frac{\min_{i\in[d]} \mathbb{E}[g_i^2]}{\varepsilon + \min_{i\in[d]} \mathbb{E}[g_i^2]} \cdot I \tag{48}$$

Algorithm 8 Diagonal RMSProp with burn-in

Input: initial x_0, time T, stepsize η, burn-in length W
$\hat{v}_0 \leftarrow \mathrm{diag}(\textsc{BurnIn}(W, \beta))$ ▷ Appendix B
for $t = 0, \ldots, T$ **do**
  $g_t \leftarrow$ stochastic gradient
  $\hat{v}_t \leftarrow \beta \hat{v}_{t-1} + (1-\beta) g_t^2$
  $\hat{A}_t \leftarrow \mathrm{diag}(\hat{v}_t)^{-1/2}$
  $x_{t+1} \leftarrow x_t - \eta \hat{A}_t g_t$
end for

Algorithm 9 Diagonal RMSProp with increasing stepsize

Input: initial x_0, time T, stepsizes η, r, threshold t_{thresh}, time S, burn-in length W, momentum β
$\hat{v}_0 \leftarrow \mathrm{diag}(\textsc{BurnIn}(W, \beta))$ ▷ Algorithm 5
for $t = 0, \ldots, T$ **do**
  $g_t \leftarrow$ stochastic gradient at x_t
  $\hat{v}_t \leftarrow \beta \hat{v}_{t-1} + (1-\beta) g_t^2$
  $A_t \leftarrow \mathrm{diag}(\hat{v}_t)^{-1/2}$
  if $t \bmod t_{\mathrm{thresh}} = 0$ **then**
    $x_{t+1} \leftarrow x_t - r A_t g_t$
    $\hat{v}_t \leftarrow \mathrm{diag}(\textsc{Hallucinate}(S, \beta, \mathrm{diag}(\hat{v}_t), x_t, x_{t+1}))$ ▷ Algorithm 6
  else
    $x_{t+1} \leftarrow x_t - \eta A_t g_t$
  end if
end for

so that overall we can take

$$\nu = \lambda_{\min}(G\,\mathrm{diag}(G)^{-1}) \cdot \frac{\min_{i \in [d]} \mathbb{E}[g_i^2]}{\varepsilon + \min_{i \in [d]} \mathbb{E}[g_i^2]} \leq \lambda_{\min}(G(\varepsilon + \mathrm{diag}(G))^{-1}). \tag{49}$$

Finally,

$$\lambda_- = \lambda_{\min}(A) = \frac{1}{(\max_{i \in [d]} \mathbb{E}[g_i^2] + \varepsilon)^{1/2}}. \tag{50}$$

□

D. Convergence results for the diagonal case

In this section we give convergence results for the diagonal approximation $A = \mathrm{diag}(\mathbb{E}[g^2] + \varepsilon)^{-1/2}$.

There are three interacting components to the results. First, using estimates $Y_t = \mathrm{diag}(g_t^2)$, Theorem 4.1 says we can accurately estimate $\mathbb{E}[Y_t] = \mathrm{diag}(\mathbb{E}[g_t^2])$ via an exponential moving average, under reasonable assumptions. Second, the curvature and noise constants for this case are already given in Appendix C, specifically Proposition C.3. Finally, we plug these results in, together with Theorems 4.3 and 4.4, to get convergence bounds for the common diagonal version of RMSProp:

Corollary D.1. *Consider diagonal RMSProp with burn-in, as in Algorithm 8, where we estimate $A = (\mathbb{E}[g^2] + \varepsilon)^{-1/2}$. Retain the same choice of $\eta = O(\tau^2)$ and $T = O(\tau^{-4})$ as in Theorem 4.3. For small enough τ, such a choice of η will yield $\Delta < \lambda_-/2$. Choose all other parameters e.g. β in accordance with Proposition 4.2. In particular, choose $W = \Theta(\eta^{-2/3}) = \Theta(\tau^{-4/3}) = O(T)$ for the burn-in parameter. Then with probability $1 - \delta$, in overall time $O(W + T) = O(\tau^{-4})$, we achieve*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \tau^2. \tag{51}$$

Corollary D.2. *Consider the diagonal RMSProp version of Algorithm 4 that is formalized in Algorithm 9. Retain the same choice of* $\eta = O(\tau^{5/2})$, $r = O(\tau)$, *and* $T = O(\tau^{-5})$ *as in Theorem 4.4. For small enough* τ, *such a choice of* η *will yield* $\Delta < \lambda_-/2$. *Choose* $W = \Theta(\eta^{-2/3}) = \Theta(\tau^{-5/3}) = O(T)$ *for the burn-in parameter Choose* $S = O(\tau^{-3/2})$, *so that as far as the estimation scheme is concerned, the stepsize is bounded by* $\max\{\eta, r/S\} = O(\tau^{5/2}) = O(\eta)$. *Then with probability* $1-\delta$, *we can reach an* $(\tau, \sqrt{\rho\tau})$*-stationary point in total time*

$$W + T = \tilde{O}\left(\frac{\Lambda_1^4\Lambda_2^4\Gamma^4}{\lambda_-^{10}\nu^4} \cdot \frac{L^3}{\rho\delta^3} \cdot \tau^{-5}\right), \tag{52}$$

where $\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-$ *are the constants describing* $A = \mathrm{diag}(\mathbb{E}[g^2] + \varepsilon)^{-1/2}$.

Note that in Algorithms 8 and 9, it is simple to implement efficient diagonal versions of the BurnIn (Algorithm 5) and Hallucinate (Algorithm 6) subroutines.

E. Main Proof

Here we will study the convergence of Algorithm 4. This is the same as Algorithm 1 except that once in a while we take a large stepsize so we may escape saddlepoints.

In order to unify our results, we prove second order convergence for general preconditioners $A(x)$. The convergence rate will depend on various properties of $A(x)$, and $A = \mathbb{E}[gg^T]^{-1/2}$ will turn out to be particularly well-behaved.

E.1. Definitions

Let ρ be the Lipschitz constant of the Hessian H, and let α be the Lipschitz constant of the preconditioner matrix $A(x)$ as a function of the current iterate x. The usual stepsize is η, while r is the occasional large stepsize that happens every t_{thresh} iterations. δ is a small probability of failure, d is the dimension. Since it will recur often, we define $\kappa = (1 + \eta\gamma)$, where γ is the magnitude of the most negative eigenvalue of $A^{1/2}HA^{1/2}$. By the following lemma, we will be able to lower bound γ by $\lambda_{\min}(A)|\lambda_{\min}(H)| \geq \lambda_-\sqrt{\rho\tau}$:

Lemma E.1. *Suppose* A *and* H *are symmetric matrices, with* $A \succ 0$ *and* $\lambda_{\min}(H) < 0$. *Then there is a negative eigenvalue of* $A^{1/2}HA^{1/2}$ *with magnitude at least* $\lambda_{\min}(A)|\lambda_{\min}(H)|$.

Proof. Let v be the minimum eigenvector of H, so that $v^THv = -\lambda_{\min}(H)\|v\|^2 = -\lambda_{\min}(H)$. Define the unit vector $u = A^{-1/2}v/\|A^{-1/2}v\|$. Then,

$$u^TA^{1/2}HA^{1/2}u = \frac{1}{\|A^{-1/2}v\|^2}v^THv = -\frac{\lambda_{\min}(H)}{\|A^{-1/2}v\|^2}. \tag{53}$$

The vector u is not necessarily an eigenvector of $A^{1/2}HA^{1/2}$, but the above expression guarantees that $A^{1/2}HA^{1/2}$ has a negative eigenvalue with magnitude at least

$$\frac{\lambda_{\min}(H)}{\|A^{-1/2}v\|^2} \geq \frac{\lambda_{\min}(H)}{\lambda_{\max}(A^{-1})\|v\|^2} = \lambda_{\min}(H)\lambda_{\min}(A). \tag{54}$$

□

Throughout, we will assume that A is a $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$-preconditioner, that $\hat{A}$ also satisfies the Λ_1 inequality, and that $\|\hat{A} - A\| \leq \Delta$.

Differing from Daneshmand et al. (2018), we will assume a uniform bound on $\|Ag\| \leq M$. In general this bound need not depend on either the spectrum of A or any uniform bound on g. For example, if g were Gaussian, Ag would be a Gaussian with zero mean and identity covariance, so we would expect $\|Ag\| = O(\sqrt{d})$ with high probability. In general M should have the same scale as $\sqrt{\Gamma}$, and the statement of Theorem 4.4 reflects this.

The proofs rely on a few other quantities that we will optimally determine as a function of the problem parameters: f_{thresh} is a threshold on the function value progress, and $g_{\text{thresh}} = f_{\text{thresh}}/t_{\text{thresh}}$ is the time-amortized average of f_{thresh}.

E.2. High level picture

For shorthand we write $A_t := A(x_t)$. Since we want to converge to a second order stationary point, our overall goal is to study the event

$$\mathcal{E}_t := \{\|\nabla f(x_t)\| \geq \tau \text{ or } \lambda_{\min}(\nabla^2 f(x_t)) \leq -\sqrt{\rho}\tau^{1/2}\} \tag{55}$$
$$= \{\|\nabla f(x_t)\| \geq \tau \text{ or } (\|\nabla f(x_t)\| \leq \tau \text{ and } \lambda_{\min}(\nabla^2 f(x_t)) \leq -\sqrt{\rho}\tau^{1/2})\}. \tag{56}$$

(where t is obvious from context, we will omit it. In words, $\mathcal{E}_t$ is the event that we are not at a second order stationary point. The main theorem results from bounding the progress we make when $\mathcal{E}_t$ does not yet hold, while also ensuring we do not leave once we hit a second order stationary point:

Lemma E.2. *Suppose that both*

$$\mathbb{E}[f(x_{t+1}) - f(x_t)|\mathcal{E}_t] \leq -g_{\text{thresh}} \tag{57}$$
$$\textit{and } \mathbb{E}[f(x_{t+1}) - f(x_t)|\mathcal{E}_t^c] \leq \delta g_{\text{thresh}}/2. \tag{58}$$

Set $T = 2(f(x_0) - \min_x f(x))/(\delta g_{\text{thresh}})$. We return x_t uniformly randomly from $x_1, \ldots, x_T$. Then, with probability at least $1 - \delta$, we will have chosen a time t where $\mathcal{E}_t$ did not occur.

Proof. Let P_t be the probability that $\mathcal{E}_t$ occurs. Then,

$$\mathbb{E}[f(x_{t+1}) - f(x_t)] = \mathbb{E}[f(x_{t+1}) - f(x_t)|\mathcal{E}_t]P_t + \mathbb{E}[f(x_{t+1}) - f(x_t)|\mathcal{E}_t^c](1 - P_t) \tag{59}$$
$$\leq -g_{\text{thresh}}P_t + \delta g_{\text{thresh}}/2 \cdot (1 - P_t) \tag{60}$$
$$\leq \delta g_{\text{thresh}}/2 - (1 + \delta/2)g_{\text{thresh}}P_t \tag{61}$$
$$\leq \delta g_{\text{thresh}}/2 - g_{\text{thresh}}P_t. \tag{62}$$

Summing over all T iterations, we have:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[f(x_{t+1}) - f(x_t)] \leq g_{\text{thresh}} \cdot \frac{1}{T}\sum_{t=1}^{T}(\delta/2 - P_t) \tag{63}$$
$$\implies \frac{1}{T}\sum_{t=1}^{T}P_t \leq \delta/2 + \frac{f(x_0) - f^*}{T g_{\text{thresh}}} \leq \delta \tag{64}$$
$$\implies \frac{1}{T}\sum_{t=1}^{T}(1 - P_t) \geq 1 - \delta. \tag{65}$$

□

Theorem E.1. *Write $\gamma = \lambda_- \sqrt{\rho}\tau^{1/2}$. Let K be a universal constant. The parameter ω will be set later and depends only logarithmically on the other parameters. Set*

$$r = \gamma^2 \cdot \frac{\delta\nu K}{54\Lambda_1\Lambda_2\Gamma L\rho M}$$
$$\eta = \gamma^5 \cdot \frac{\delta^2\nu^2K^2}{324M^2L^2\Lambda_1^2\Lambda_2^2\Gamma^2\rho^2\omega}$$
$$f_{\text{thresh}} = \gamma^4 \cdot \frac{\delta\nu^2K^2}{54 \cdot 12\Lambda_1^2\Lambda_2^2\Gamma L\rho^2M^2}.$$

Let $t_{\text{thresh}} = \omega/(\eta\gamma)$, $\Delta = O(\tau^{1/2})$, and set $g_{\text{thresh}} = f_{\text{thresh}}/t_{\text{thresh}}$. Then we have both

$$\mathbb{E}[f(x_{t+1}) - f(x_t)|\mathcal{E}_t] \leq -g_{\text{thresh}} \tag{66}$$
$$\textit{and } \mathbb{E}[f(x_{t+1}) - f(x_t)|\mathcal{E}_t^c] \leq \delta g_{\text{thresh}}/2. \tag{67}$$

Corollary E.1. *In the above setting, with probability $1-\delta$, we reach an $(\tau, \sqrt{\rho}\tau^{1/2})$-stationary point in time*

$$\tilde{O}\left(\frac{M^4 L^3}{\rho\delta^3} \cdot \frac{\Lambda_1^4 \Lambda_2^4 \Gamma^2}{\lambda_-^{10}\nu^4} \cdot \tau^{-5}\right). \tag{68}$$

Proof. Simply observe $T = C(f_0 - f^*)/(\delta g_{\mathrm{thresh}})$ and plug in g_{thresh}. □

E.3. Amortized increase due to large stepsize iterations

Before we start casework on whether $\mathcal{E}_t$ holds We want to bound the amortized effect on the objective of increasing the stepsize every t_{thresh} iterations. By Corollary E.2,

$$\mathbb{E}[f(x_{t+1})] - f(x_t) \le \frac{9L\Gamma r^2}{8}. \tag{69}$$

Note that for our particular setting of r and f_{thresh}, we have

$$\frac{9L\Gamma}{8} \cdot r^2 = \frac{9L\Gamma}{8} \cdot \gamma^4 \cdot \frac{\delta^2\nu^2 K^2}{54^2 \Lambda_1^2 \Lambda_2^2 \Gamma^2 L^2 \rho^2 M^2} \tag{70}$$

$$= \frac{9\delta}{8} \cdot \frac{12}{54} \cdot \gamma^4 \cdot \frac{\delta\nu^2 K^2}{54 \cdot 12 \Lambda_1^2 \Lambda_2^2 \Gamma L \rho^2 M^2} \tag{71}$$

$$= \frac{\delta f_{\mathrm{thresh}}}{4}, \tag{72}$$

so also

$$\mathbb{E}[f(x_{t+1})] - f(x_t) \le \frac{\delta f_{\mathrm{thresh}}}{4}. \tag{73}$$

Therefore on average

$$\frac{\mathbb{E}[f(x_{t+1})] - f(x_t)}{t_{\mathrm{thresh}}} \le \delta g_{\mathrm{thresh}}/4. \tag{74}$$

E.4. Bound on possible increase when $\mathcal{E}_t^c$ occurs

For the main result we need to bound

$$\mathbb{E}[f(x_{t+1}) - f(x_t) | \mathcal{E}_t^c] \le \delta g_{\mathrm{thresh}}/4. \tag{75}$$

Note that

$$\mathcal{E}_t^c = \{\|\nabla f(x_t)\| \ge \tau \text{ or } \lambda_{\min}(\nabla^2 f(x_t)) \le -\sqrt{\rho}\tau^{1/2}\}^c \tag{76}$$

$$= \{\|\nabla f(x_t)\| < \tau \text{ and } \lambda_{\min}(\nabla^2 f(x_t)) > -\sqrt{\rho}\tau^{1/2}\}. \tag{77}$$

Hence it suffices to bound the function increase conditioned on $\|\nabla f(x_t)\| \le \tau$. By Corollary E.2 we have

$$\mathbb{E}[f(x_{t+1})] - f(x_t) \le \frac{9L\Gamma\eta^2}{8}. \tag{78}$$

We want this to not exceed $\delta g_{\mathrm{thresh}}/4$:

$$\frac{9L\Gamma\eta^2}{8} \overset{?}{\le} \frac{\delta}{4} g_{\mathrm{thresh}} \tag{79}$$

$$\Leftrightarrow \frac{9L\Gamma\eta^2}{8} \overset{?}{\le} \frac{\delta}{4} f_{\mathrm{thresh}} \cdot \frac{\eta\gamma}{\omega} \tag{80}$$

$$\Leftrightarrow \frac{9L\Gamma\eta}{2} \overset{?}{\le} \delta f_{\mathrm{thresh}} \cdot \frac{\gamma}{\omega} \tag{81}$$

$$\Leftrightarrow \frac{9L\Gamma}{2} \cdot \gamma^5 \cdot \frac{\delta^2\nu^2 K^2}{324 M^2 L^2 \Lambda_1^2 \Lambda_2^2 \Gamma^2 \rho^2 \omega} \overset{?}{\le} \delta \cdot \frac{\gamma}{\omega} \cdot \gamma^4 \cdot \frac{\delta\nu^2 K^2}{54 \cdot 12 \Lambda_1^2 \Lambda_2^2 \Gamma L \rho^2 M^2}. \tag{82}$$

Cancelling like terms, we find that the inequality is equivalent to $\omega \geq 9/4$, which we can easily enforce later. Therefore we may indeed write that

$$\mathbb{E}[f(x_{t+1})] - f(x_t) \leq \frac{\delta g_{\text{thresh}}}{4}. \tag{83}$$

E.5. Bound on decrease (progress) when $\mathcal{E}_t$ occurs

We need to bound

$$\mathbb{E}[f(x_{t+1}) - f(x_t)|\mathcal{E}_t] \leq -g_{\text{thresh}}. \tag{84}$$

By definition,

$$\mathcal{E}_t = \{\|\nabla f(x_t)\| \geq \tau\} \cup \{\lambda_{\min}(\nabla^2 f(x_t)) \leq -\sqrt{\rho}\tau^{1/2} \text{ and } \|\nabla f(x_t)\| \leq \tau\}. \tag{85}$$

In words, we split $\mathcal{E}_t$ into two cases: either the gradient is large, or we are near a saddlepoint but there is an escape direction.

E.5.1. Large gradient regime

If the norm of the gradient is large enough, i.e.

$$\|\nabla f(x_t)\|^2 \geq \tau^2 \tag{86}$$

then by Corollary E.4,

$$\mathbb{E}[f(x_{t+1})] - f(x_t) \leq -\frac{\eta\tau^2\lambda_-}{4} \leq -g_{\text{thresh}} \tag{87}$$

as long as $\eta \leq \frac{4\lambda_-\tau^2}{9L\Gamma}$ and $g_{\text{thresh}} \leq \frac{\eta\tau^2\lambda_-}{4}$. For our choice of $\eta = O(\tau^{5/2})$ and $g_{\text{thresh}} = \tilde{O}(\tau^5)$, each of these will hold for small enough τ.

E.5.2. Sharp negative curvature regime

We start at a point x_0 around which we base our Hessian approximation:

$$g(x) = f(x_0) + (x - x_0)^T\nabla f(x_0) + \frac{1}{2}(x - x_0)^T H(x - x_0) \tag{88}$$

where we write $H = \nabla^2 f(x_0)$. We will also write $A = \mathbb{E}[g_0 g_0^T]^{1/2}$ as the preconditioner at x_0.

Lemma E.3. *For every twice differentiable ρ-Hessian Lipschitz function f we have*

$$\|\nabla f(x) - \nabla g(x)\| \leq \frac{\rho}{2}\|x - x_0\|^2. \tag{89}$$

Proof. From Lemma 1.2.4 in (Nesterov, 2004), we have

$$\|\nabla f(x) - \nabla f(x_0) - H(x - x_0)\| \leq \frac{\rho}{2}\|x - x_0\|^2. \tag{90}$$

Now simply observe that $\nabla g(x) = \nabla f(x_0) + H(x - x_0)$. □

Lemma E.4. *Suppose that $\|\nabla f(x_0)\| \leq \tau$. Also suppose the Hessian at x_0 has a strong escape direction, i.e. $\lambda_{\min}(\nabla^2 f(x_0)) \leq -\sqrt{\rho}\tau^{1/2}$, and define $\gamma = \lambda_-\sqrt{\rho}\tau^{1/2}$ so that $\sqrt{\rho}\tau^{1/2} = \lambda_-^{-1}\gamma$. Then there exists $k < t_{\text{thresh}}$ so that*

$$\mathbb{E}[f(x_k)] - f(x_0) \leq -f_{\text{thresh}} \tag{91}$$

Proof. Suppose not, i.e. suppose that for all $t < t_{\text{thresh}}$ it holds that

$$\mathbb{E}[f(x_t)] - f(x_0) \geq -f_{\text{thresh}}. \tag{92}$$

Under this assumption we will prove bounds which will imply that the assumption cannot hold. In particular, we will give a lower bound on $\mathbb{E}[\|x_t - x_0\|^2]$ that conflicts with Lemma E.11.

Define the following terms, each of which is selected to satisfy a certain recursion:

Term	Recursion identity
$u_t = (I - \eta AH)^t(x_1 - x_0)$.	$u_t = (I - \eta AH)u_{t-1}$
$\delta_t = \sum_{i=1}^t (I - \eta AH)^{t-i} A(-\nabla f(x_i) + \nabla g(x_i))$	$\delta_t = A(-\nabla f(x_t) + \nabla g(x_t)) + (I - \eta AH)\delta_{t-1}$
$d_t = -\sum_{i=1}^t (I - \eta AH)^{t-i} A\nabla f(x_0)$	$d_t = -A\nabla f(x_0) + (I - \eta AH)d_{t-1}$
$\zeta_t = \sum_{i=1}^t (I - \eta AH)^{t-i}\xi_i$	$\zeta_t = \xi_t + (I - \eta AH)\zeta_{t-1}$
$\chi_t = \sum_{i=1}^t (I - \eta AH)^{t-i}(A - A_i)\nabla f(x_i)$	$\chi_t = (A - A_t)\nabla f(x_t) + (I - \eta AH)\chi_{t-1}$
$\iota_t = \sum_{i=1}^t (I - \eta AH)^{t-i}(A_i - \hat{A}_i)\nabla f(x_i)$	$\iota_t = (A_t - \hat{A}_t)\nabla f(x_t) + (I - \eta AH)\iota_{t-1}$.

By convention we take $\delta_0 = d_0 = \zeta_0 = \chi_0 = \iota_0 = 0$, and for convenience we will write $\pi_t = \delta_t + d_t + \zeta_t + \chi_t + \iota_t$. These terms are chosen so that each is a kind of error term in a stale Taylor expansion of f. The error terms cancel so that the following identity holds:

$$\pi_t = (I - \eta AH)\pi_{t-1} + AH(x_t - x_0) - \hat{A}_t\nabla f(x_t) + \xi_t. \tag{93}$$

We will inductively prove the identity

$$\begin{aligned} x_{t+1} - x_0 &= u_t + \eta\pi_t && (94)\\ &= u_t + \eta(\delta_t + d_t + \zeta_t + \chi_t + \iota_t) && (95)\end{aligned}$$

for $t \geq 0$. The base case is simple because $u_0 = x_1 - x_0$ and the other terms are all zero. For the inductive step, note

$$\begin{aligned}(x_{t+1} - x_0) - (x_t - x_0) &= x_{t+1} - x_t\\ &= -\eta\hat{A}_t\nabla f(x_t) + \eta\xi_t\\ &= \eta[\pi_t - (I - \eta AH)\pi_{t-1} - AH(x_t - x_0)]\end{aligned}$$

where the last equality follows from equation (93). Rearranging, we have

$$\begin{aligned} x_{t+1} - x_0 &= \eta[\pi_t - (I - \eta AH)\pi_{t-1}] + (I - \eta AH)(x_t - x_0) && (96)\\ &= \eta[\pi_t - (I - \eta AH)\pi_{t-1}] + (I - \eta AH)(u_{t-1} + \eta\pi_{t-1}) && (97)\end{aligned}$$

by induction. Since $u_t = (I - \eta AH)u_{t-1}$, we can further simplify:

$$\begin{aligned} x_{t+1} - x_0 &= \eta[\pi_t - (I - \eta AH)\pi_{t-1}] + (I - \eta AH)(u_{t-1} + \eta\pi_{t-1}) && (98)\\ &= \eta[\pi_t - (I - \eta AH)\pi_{t-1}] + u_t + \eta(I - \eta AH)\pi_{t-1}) && (99)\\ &= u_t + \eta\pi_t && (100)\end{aligned}$$

as desired.

To proceed with our saddle point escape argument, we must bound all the terms $u_t, \delta_t, d_t, \zeta_t, \chi_t, \iota_t$ to show that $x_t - x_0$ grows fast enough.

Lemma E.5. *Under the above conditions, we have*

$$\mathbb{E}[\|\chi_t\|] \leq \alpha\tau\sqrt{\eta^3 L\Gamma\Lambda_1} \cdot \kappa^t \cdot \left(\frac{4}{(\eta\gamma)^2} + \frac{6 f_{\text{thresh}}}{\eta^3\gamma L\Gamma} + \frac{2}{\eta\gamma} \cdot \sqrt{\frac{2r^2}{\eta^3 L\Lambda_1}}\right). \tag{101}$$

Proof. We assume $A(x)$ is α Lipschitz, so that $\|A_i - A\| \leq \alpha \|x_i - x_0\|$. Then,

$$\mathbb{E}[\|\chi_t\|] = \mathbb{E}\left[\left\|\sum_{i=1}^{t}(I-\eta AH)^{t-i}(A-A_i)\nabla f(x_i)\right\|\right] \tag{102}$$

$$\leq \sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\,\mathbb{E}\left[\|(A-A_i)\nabla f(x_i)\|\right] \tag{103}$$

$$\leq \sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\,\mathbb{E}\left[\|A-A_i\|\|\nabla f(x_i)\|\right] \tag{104}$$

$$\leq \tau\sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\,\mathbb{E}\left[\|A-A_i\|\right] \tag{105}$$

$$\leq \alpha\tau\sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\,\mathbb{E}\left[\|x_i-x_0\|\right] \tag{106}$$

$$\leq \alpha\tau\sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\sqrt{\mathbb{E}\left[\|x_i-x_0\|^2\right]} \tag{107}$$

$$\leq \alpha\tau\sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\sqrt{6\eta f_{\text{thresh}}\Lambda_1 i + \eta^3 L\Gamma\Lambda_1 i^2 + 2\Gamma r^2} \tag{108}$$

where for the last identity we have applied Lemma E.11. By Lemma E.12, we may further bound this by

$$\mathbb{E}[\|\chi_t\|] \leq \alpha\tau\sqrt{\eta^3 L\Gamma\Lambda_1}\sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\left(2i + \frac{3\eta f_{\text{thresh}}\Lambda_1}{\eta^3 L\Gamma\Lambda_1} + \sqrt{\frac{2\Gamma r^2}{\eta^3 L\Gamma\Lambda_1}}\right) \tag{109}$$

$$= \alpha\tau\sqrt{\eta^3 L\Gamma\Lambda_1}\sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\left(2i + \frac{3 f_{\text{thresh}}}{\eta^2 L\Gamma} + \sqrt{\frac{2r^2}{\eta^3 L\Lambda_1}}\right). \tag{110}$$

Applying Lemma E.14 with $\beta = \eta\gamma$ yields:

$$\mathbb{E}[\|\chi\|] \leq \alpha\tau\sqrt{\eta^3 L\Gamma\Lambda_1}\cdot\kappa^t\cdot\left(\frac{4}{(\eta\gamma)^2} + \frac{2}{\eta\gamma}\cdot\frac{3f_{\text{thresh}}}{\eta^2 L\Gamma} + \frac{2}{\eta\gamma}\cdot\sqrt{\frac{2r^2}{\eta^3 L\Lambda_1}}\right) \tag{111}$$

$$= \alpha\tau\sqrt{\eta^3 L\Gamma\Lambda_1}\cdot\kappa^t\cdot\left(\frac{4}{(\eta\gamma)^2} + \frac{6f_{\text{thresh}}}{\eta^3\gamma L\Gamma} + \frac{2}{\eta\gamma}\cdot\sqrt{\frac{2r^2}{\eta^3 L\Lambda_1}}\right). \tag{112}$$

□

Lemma E.6. *Under the above conditions, we have*

$$\mathbb{E}[\|\delta_t\|] \leq \Lambda_2\rho\kappa^t\left[\frac{2\Gamma r^2}{\eta\gamma} + \frac{6\eta f_{\text{thresh}}\Lambda_1}{(\eta\gamma)^2} + \frac{3\eta^3 L\Gamma\Lambda_1}{(\eta\gamma)^3}\right]. \tag{113}$$

Proof. We write

$$\mathbb{E}[\|\delta_t\|] = \mathbb{E}\left[\left\|\sum_{i=1}^{t}(I-\eta AH)^{t-i}A(\nabla f(x_i)-\nabla g(x_i))\right\|\right] \tag{114}$$

$$\leq \sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\,\mathbb{E}\left[\|A(\nabla f(x_i)-\nabla g(x_i))\|\right] \tag{115}$$

$$\leq \Lambda_2\sum_{i=1}^{t}\kappa^{t-i}\,\mathbb{E}\left[\|\nabla f(x_i)-\nabla g(x_i)\|\right] \tag{116}$$

$$\leq \Lambda_2(\rho/2)\sum_{i=1}^{t}\kappa^{t-i}\,\mathbb{E}\left[\|x_i-x_0\|^2\right] \tag{117}$$

$$\leq \Lambda_2(\rho/2)\sum_{i=1}^{t}\kappa^{t-i}\left(6\eta f_{\text{thresh}}\Lambda_1 i+\eta^3 L\Gamma\Lambda_1 i^2+2\Gamma r^2\right), \tag{118}$$

where again, the last inequality comes from Lemma E.11. Applying Lemma E.14 with $\beta = \eta\gamma$ yields:

$$\mathbb{E}[\|\delta_t\|] \leq \frac{\Lambda_2\rho\kappa^t}{2}\left[(6\eta f_{\text{thresh}}\Lambda_1)\cdot\frac{2}{\eta^2\gamma^2}+\eta^3 L\Gamma\Lambda_1\cdot\frac{6}{\eta^3\gamma^3}+2\Gamma r^2\cdot\frac{2}{\eta\gamma}\right] \tag{119}$$

$$= \Lambda_2\rho\kappa^t\left[\frac{2\Gamma r^2}{\eta\gamma}+\frac{6\eta f_{\text{thresh}}\Lambda_1}{(\eta\gamma)^2}+\frac{3\eta^3 L\Gamma\Lambda_1}{(\eta\gamma)^3}\right]. \tag{120}$$

□

Lemma E.7. *Under the above conditions,*

$$\mathbb{E}\|\iota_t\|] \leq 2\tau(\eta\gamma)^{-1}\Delta\kappa^t. \tag{121}$$

Proof. Write

$$\mathbb{E}\|\iota_t\|] = \mathbb{E}\left[\left\|\sum_{i=1}^{t}(I-\eta AH)^{t-i}(A_i-\hat{A}_i)\nabla f(x_i)\right\|\right] \tag{122}$$

$$\leq \sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\,\mathbb{E}\left[\left\|(A_i-\hat{A}_i)\nabla f(x_i)\right\|\right] \tag{123}$$

$$\leq \tau\sum_{i=1}^{t}(1+\eta\gamma)^{t-i}\,\mathbb{E}\left[\left\|A_i-\hat{A}_i\right\|\right] \tag{124}$$

$$\leq 2\tau(\eta\gamma)^{-1}\kappa^t\max_i\mathbb{E}\left[\left\|A_i-\hat{A}_i\right\|\right] \tag{125}$$

$$\leq 2\tau(\eta\gamma)^{-1}\Delta\kappa^t. \tag{126}$$

□

Lemma E.8. *Under the above conditions,* $\mathbb{E}[u_t^T]d_t \geq 0$.

Proof. We have

$$\mathbb{E}[u_t] = (I-\eta AH)^t\,\mathbb{E}[x_1-x_0] = -r(I-\eta AH)^t A\nabla f(x_0). \tag{127}$$

For small enough η, we have $\|\eta AH\| \leq 1$ and hence:

$$\mathbb{E}[u_t^T]d_t = r\left[(I-\eta AH)^t A\nabla f(x_0)\right]^T\sum_{i=1}^{t}(I-\eta AH)^{t-i}A\nabla f(x_0) \tag{128}$$

$$= r\sum_{i=1}^{t}(A\nabla f(x_0))^T(I-\eta AH)^{2t-i}(A\nabla f(x_0)) \geq 0. \tag{129}$$

□

Lemma E.9. *Under the above conditions, we get an exponentially growing lower bound on the expected squared norm of u_t:*

$$\mathbb{E}[\|u_t\|^2] \geq (1+\eta\gamma)^{2t} r^2 \nu = \kappa^{2t} r^2 \nu. \tag{130}$$

Proof. For unit vectors v, we may write

$$\mathbb{E}[\|u_t\|^2] \geq \mathbb{E}[(v^T u_t)^2]. \tag{131}$$

In particular, by definition of u_t,

$$\mathbb{E}[\|u_t\|^2] \geq \mathbb{E}[(v^T (I - \eta AH)^t (x_1 - x_0))^2]. \tag{132}$$

We wish to choose a unit vector v so that this is as large as possible. If AH were symmetric, we could choose v to be an eigenvector, but the product of symmetric matrices is not in general symmetric. However, because A and H are both symmetric, and A is positive definite, it follows that $A^{1/2}$ exists and that $A^{1/2}HA^{1/2}$ is symmetric. Hence for orthonormal U and diagonal Λ, we have

$$A^{1/2}HA^{1/2} = U\Lambda U^T \tag{133}$$

$$\implies A^{-1/2}AHA^{1/2} = U\Lambda U^T \tag{134}$$

$$\implies AH = A^{1/2}U\Lambda(A^{1/2}U)^{-1}. \tag{135}$$

The diagonal matrix Λ contains the eigenvalues of $A^{1/2}HA^{1/2}$. Without loss of generality, Λ_{11} corresponds to a negative eigenvalue with absolute value γ. Therefore

$$(I - \eta AH)^t = (A^{1/2}U(I - \eta\Lambda)(A^{1/2}U)^{-1})^t \tag{136}$$

$$= A^{1/2}U(I - \eta\Lambda)^t(A^{1/2}U)^{-1}. \tag{137}$$

Since we can choose v to be any unit vector we want, we will set it equal to $C(U^T A^{1/2})^{-1}e_1$ so that $U^T A^{1/2}v = Ce_1$. Here e_1 is the first standard basis vector and C is a scalar constant chosen to make v a unit vector. Taking transposes, we have $v^T A^{1/2}U = Ce_1^T$. Now,

$$v^T(I - \eta AH)^t = v^T A^{1/2}U(I - \eta\Lambda)^t(A^{1/2}U)^{-1} \tag{138}$$

$$= Ce_1^T(I - \eta\Lambda)^t(A^{1/2}U)^{-1} \tag{139}$$

$$= C(1 + \eta\Lambda_{11})^t e_1^T(A^{1/2}U)^{-1} \tag{140}$$

$$= (1 + \eta\gamma)^t \cdot Ce_1^T(A^{1/2}U)^{-1}. \tag{141}$$

Substituting in the definition of v, this is equal to:

$$v^T(I - \eta AH)^t = (1 + \eta\gamma)^t \cdot v^T(A^{1/2}U)(A^{1/2}U)^{-1} \tag{142}$$

$$= (1 + \eta\gamma)^t v^T. \tag{143}$$

This equality holds for any v of the form specified above; in particular, choose C so that v is unit. Then, we may finally bound

$$\mathbb{E}[\|u_t\|^2] \geq \mathbb{E}[(v^T(I - \eta AH)^t(x_1 - x_0))^2] \tag{144}$$

$$\geq (1 + \eta\gamma)^{2t}\,\mathbb{E}[(v^T(x_1 - x_0))^2] \tag{145}$$

$$= (1 + \eta\gamma)^{2t} r^2\,\mathbb{E}[(v^T A g_0)^2] \tag{146}$$

$$= (1 + \eta\gamma)^{2t} r^2 v^T\,\mathbb{E}[A g_0 g_0^T A^T] v \tag{147}$$

$$= (1 + \eta\gamma)^{2t} r^2 v^T A\,\mathbb{E}[g_0 g_0^T] A^T v \tag{148}$$

$$\geq (1 + \eta\gamma)^{2t} r^2 \lambda_{\min}(A\,\mathbb{E}[g_0 g_0^T] A^T) \tag{149}$$

$$\geq (1 + \eta\gamma)^{2t} r^2 \nu, \tag{150}$$

where the last two lines follow by the fact that $\|v\| = 1$ and by definition of ν. □

Lemma E.10. *Under the above conditions we have a deterministic bound on $\|u_t\|$:*

$$\|u_t\| \leq \kappa^t r M \tag{151}$$

Proof. We write

$$\begin{aligned}
\|u_t\| &= \|(I - \eta A H)^t (x_1 - x_0)\| &(152)\\
&\leq \|I - \eta A H\|^t \cdot \|x_1 - x_0\| &(153)\\
&\leq (1 + \eta\gamma)^t \cdot r \|A g_0\| &(154)\\
&\leq (1 + \eta\gamma)^t \cdot r M. &(155)
\end{aligned}$$

□

Putting all these results together, we can give a lower bound on the distance between iterates:

$$\begin{aligned}
\mathbb{E}[\|x_{t+1} - x_0\|^2] &= \mathbb{E}\left[\|u_t + \eta(\delta_t + d_t + \zeta_t + \chi_t + \iota_t)\|^2\right] \\
&= \mathbb{E}[\|u_t\|^2] + 2\eta\,\mathbb{E}\left[u_t^T(\delta_t + d_t + \zeta_t + \chi_t + \iota_t)\right] + \eta^2\,\mathbb{E}\left[\|\delta_t + d_t + \zeta_t + \chi_t + \iota_t\|^2\right] \\
&\geq \mathbb{E}[\|u_t\|^2] + 2\eta\,\mathbb{E}\left[u_t^T(\delta_t + d_t + \zeta_t + \chi_t + \iota_t)\right] \\
&= \mathbb{E}[\|u_t\|^2] + 2\eta\,\mathbb{E}\left[u_t^T(\delta_t + d_t + \chi_t + \iota_t)\right] \\
&= \mathbb{E}[\|u_t\|^2] + 2\eta\,\mathbb{E}[u_t^T\delta_t] + 2\eta\,\mathbb{E}[u_t^T d_t] + 2\eta\,\mathbb{E}[u_t^T\chi_t] \\
&= \mathbb{E}[\|u_t\|^2] + 2\eta\,\mathbb{E}[u_t^T\delta_t] + 2\eta\,\mathbb{E}[u_t]^T d_t + 2\eta\,\mathbb{E}[u_t^T\chi_t] + 2\eta\,\mathbb{E}[u_t^T\iota_t] \\
&\geq \mathbb{E}[\|u_t\|^2] + 2\eta\,\mathbb{E}[u_t^T\delta_t] + 2\eta\,\mathbb{E}[u_t^T\chi_t] + 2\eta\,\mathbb{E}[u_t^T\iota_t] \\
&\geq \mathbb{E}[\|u_t\|^2] - 2\eta\|u_t\|\,\mathbb{E}[\|\delta_t\|] - 2\eta\|u_t\|\,\mathbb{E}[\|\chi_t\|] - 2\eta\|u_t\|\,\mathbb{E}[\|\iota_t\|] \\
&\geq \kappa^{2t} r^2 \nu - 2\eta\kappa^t r M\,\mathbb{E}[\|\delta_t\| + \|\chi_t\| + \|\iota_t\|].
\end{aligned}$$

Substituting in the bounds for $\mathbb{E}[\|\delta_t\|]$, $\mathbb{E}[\|\chi_t\|]$, and $\mathbb{E}[\|\iota_t\|]$,we finally have the lower bound:

$$\begin{aligned}
&\left(r\nu - 2\eta M\left[\Lambda_2\rho\left[\frac{2\Gamma r^2}{\eta\gamma} + \frac{6\eta f_{\text{thresh}}\Lambda_1}{(\eta\gamma)^2} + \frac{3\eta^3 L\Gamma\Lambda_1}{(\eta\gamma)^3}\right]\right.\right. &(156)\\
&\left.\left.+\alpha\tau\sqrt{\eta^3 L\Gamma\Lambda_1}\left(\frac{4}{(\eta\gamma)^2} + \frac{6 f_{\text{thresh}}}{\eta^3\gamma L\Gamma} + \frac{2}{\eta\gamma}\cdot\sqrt{\frac{2r^2}{\eta^3 L\Lambda_1}}\right) + 2\tau(\eta\gamma)^{-1}\Delta\right]\right) r\kappa^{2t}. &(157)
\end{aligned}$$

As long as the sum in the parentheses is positive, this term will grow exponentially and grant us the contradiction we seek. We want to bound each of the seven terms in brackets by $r\nu/8$, so that the overall bound is $r^2\kappa^{2t}\nu/8$. For simplicity, we will write $K = 1/8$ as a universal constant. Then, we want to choose parameters so the following inequalities all hold.

We start with the last term (from ι_t) because it is the most simple. Since $\gamma = \Theta(\tau^{1/2})$, we require that

$$\begin{aligned}
2\eta M \cdot 2\tau(\eta\gamma)^{-1}\Delta &\leq r\nu K &(158)\\
\Leftrightarrow 4M\tau\gamma^{-1}\Delta &\leq r\nu K &(159)\\
\Leftrightarrow \tau\cdot\tau^{-1/2}\Delta &\leq O(r) &(160)\\
\Leftrightarrow \Delta &\leq O(\tau^{-1/2} r). &(161)
\end{aligned}$$

Since we will eventually set $r = O(\tau)$, this constraint is simply $\Delta \leq O(\tau^{1/2})$.

Next we move onto the first three terms, which correspond to δ_t:

$$2\eta M\Lambda_2\rho\cdot\frac{2\Gamma r^2}{\eta\gamma} \leq r\nu K \Leftrightarrow r \leq \frac{\gamma\nu K}{4\Lambda_2\Gamma\rho M} \tag{162}$$

$$2\eta M\Lambda_2\rho\cdot\frac{6\eta f_{\text{thresh}}\Lambda_1}{\eta^2\gamma^2} \leq r\nu K \Leftrightarrow f_{\text{thresh}} \leq \frac{\gamma^2 r\nu K}{12\Lambda_1\Lambda_2\rho M} \tag{163}$$

$$2\eta M\Lambda_2\rho\cdot\frac{3\eta^3 L\Gamma\Lambda_1}{\eta^3\gamma^3} \leq r\nu K \Leftrightarrow \eta \leq \frac{\gamma^3 r\nu K}{6ML\Lambda_1\Lambda_2\Gamma\rho}. \tag{164}$$

The first constraint is satisfied for small enough τ because we chose $r = O(\tau) \le O(\tau^{1/2})$. The second term is equivalent to

$$f_{\text{thresh}} \overset{?}{\le} \frac{\gamma^2 \nu K}{12\Lambda_1\Lambda_2\rho M} \cdot r \tag{165}$$

$$\Leftrightarrow \gamma^4 \cdot \frac{\delta\nu^2 K^2}{54 \cdot 12\Lambda_1^2\Lambda_2^2\Gamma L\rho^2 M^2} \overset{?}{\le} \frac{\gamma^2\nu K}{12\Lambda_1\Lambda_2\rho M} \cdot \gamma^2 \cdot \frac{\delta\nu K}{54\Lambda_1\Lambda_2\Gamma L\rho M} \tag{166}$$

$$\Leftrightarrow \frac{\delta\nu^2 K^2}{54 \cdot 12\Lambda_1^2\Lambda_2^2\Gamma L\rho^2 M^2} \overset{?}{\le} \frac{\delta\nu^2 K^2}{54 \cdot 12\Lambda_1^2\Lambda_2^2\Gamma L\rho^2 M^2} \tag{167}$$

which trivially always holds since the two expressions are equal.

Finally, we address the three terms corresponding to χ_t. For small enough τ, it will turn out that none of the resulting constraints are tight, i.e. they are all weaker than some other constraint we already require. First,

$$2\eta M\alpha\tau\sqrt{\eta^3 L\Gamma\Lambda_1} \cdot \frac{4}{\eta^2\gamma^2} \le r\nu K \tag{168}$$

$$\Leftarrow \eta^{1/2}\tau \le O(r\gamma^2) \tag{169}$$

$$\Leftrightarrow \eta \le O(r^2\gamma^4\tau^{-2}) = O(\tau^2). \tag{170}$$

Next,

$$2\eta M\alpha\tau\sqrt{\eta^3 L\Gamma\Lambda_1} \cdot \frac{6f_{\text{thresh}}}{\eta^3\gamma L\Gamma} \le r\nu K \tag{171}$$

$$\Leftarrow \eta\tau\eta^{3/2}\frac{f_{\text{thresh}}}{\eta^3\gamma} \le O(r) \tag{172}$$

$$\Leftrightarrow f_{\text{thresh}} \le O(\eta^{1/2} r\gamma\tau^{-1}) = O(\tau^{7/4}). \tag{173}$$

Finally,

$$2\eta M\alpha\tau\sqrt{\eta^3 L\Gamma\Lambda_1} \cdot \frac{2}{\eta\gamma} \cdot \sqrt{\frac{2r^2}{\eta^3 L\Lambda_1}} \le r\nu K \tag{174}$$

$$\Leftarrow \tau\sqrt{\eta^3} \cdot \frac{1}{\gamma} \cdot \frac{r}{\sqrt{\eta^3}} \le O(r) \tag{175}$$

$$\Leftrightarrow \tau\gamma^{-1} r \le O(r) \tag{176}$$

$$\Leftrightarrow \tau \le O(\gamma) = O(\tau^{1/2}). \tag{177}$$

Hence, for small enough τ, for the above parameter settings, we have

$$\mathbb{E}[\|x_{t+1} - x_0\|^2] \ge r^2\kappa^{2t}\nu K. \tag{178}$$

We now have a lower bound and an upper bound that when combined yield $(1+\eta\gamma)^{2t} \le C$, where

$$C = \left[(6\eta f_{\text{thresh}}\Lambda_1)\, t + \eta^3 L\Gamma\Lambda_1 t^2 + 2\Gamma r^2\right] \cdot \frac{1}{r^2\nu K}. \tag{179}$$

We can choose ω that is only logarithmic in all parameters, i.e. $\omega = O(\log(\frac{\Lambda_1\Lambda_2\Gamma L\eta f_{\text{thresh}}}{\nu r}))$, so that setting $t \ge t_{\text{thresh}} = \omega/(\eta\gamma)$ yields $(1+\eta\gamma)^{2t} \ge C$. This contradicts the upper bound, as desired.

□

Lemma E.11. *Assume that Equation (92) holds. Assume also that $\eta \le \frac{f_{\text{thresh}}\Lambda_1}{\Gamma}$. Then,*

$$\mathbb{E}[\|x_t - x_0\|^2] \le 6\eta f_{\text{thresh}}\Lambda_1 t + \eta^3 L\Gamma\Lambda_1 t^2 + 2\Gamma r^2. \tag{180}$$

Proof. By Lemma E.16,

$$-f_{\text{thresh}} \le \mathbb{E}[f(x_t)] - f(x_0) \tag{181}$$

$$= \mathbb{E}\left[\sum_{i=0}^{t-1} f(x_{i+1}) - f(x_i)\right] \tag{182}$$

$$\le -\eta \sum_{i=0}^{t-1} \mathbb{E}[\|\hat{A}_i^{1/2} \nabla f(x_i)\|^2] + \frac{\eta^2 L\Gamma(t-1)}{2} + \frac{r^2 L\Gamma}{2}. \tag{183}$$

Remember, we are making the simplifying assumption that Λ_1 serves as a bound in the same way for $\hat{A}$ as it does for A. This is trivially true if $\Delta = 0$. Applying the definition of Λ_1 yields:

$$-f_{\text{thresh}} \le -\eta \Lambda_1^{-1} \sum_{i=0}^{t-1} \mathbb{E}[\|\hat{A}_i \nabla f(x_i)\|^2] + \frac{\eta^2 L\Gamma t}{2} + \frac{r^2 L\Gamma}{2}. \tag{184}$$

By rearranging, we can get a bound on the gradient norms:

$$\sum_{i=0}^{t-1} \mathbb{E}[\|\hat{A}_i \nabla f(x_i)\|^2] \le \frac{\Lambda_1}{\eta}\left(\frac{\eta^2 L\Gamma t}{2} + \frac{r^2 L\Gamma}{2} + f_{\text{thresh}}\right) \tag{185}$$

$$= \frac{\eta L\Gamma\Lambda_1 t}{2} + \frac{r^2 L\Gamma\Lambda_1}{2\eta} + \frac{f_{\text{thresh}}\Lambda_1}{\eta}. \tag{186}$$

Before we proceed, note that we already have

$$\frac{\delta f_{\text{thresh}}}{4} \ge \frac{9L\Gamma r^2}{8} \implies \frac{f_{\text{thresh}}\Lambda_1}{\eta} \ge \frac{9}{2\delta}\frac{r^2 L\Gamma\Lambda_1}{\eta} \ge \frac{r^2 L\Gamma\Lambda_1}{2\eta}. \tag{187}$$

Hence we can further bound equation (186) by

$$\sum_{i=0}^{t-1} \mathbb{E}[\|\hat{A}_i \nabla f(x_i)\|^2] \le \frac{\eta L\Gamma\Lambda_1 t}{2} + \frac{2 f_{\text{thresh}}\Lambda_1}{\eta}. \tag{188}$$

Now we will work toward bounding the norm of the difference $x_t - x_0$. We will first bound the difference $x_t - x_1$, then the difference $x_1 - x_0$.

$$\mathbb{E}[\|x_t - x_1\|^2] \le \mathbb{E}\left[\left\|\sum_{i=1}^{t-1} x_{i+1} - x_i\right\|^2\right] \tag{189}$$

$$\le \eta^2\, \mathbb{E}\left[\left\|\sum_{i=1}^{t-1}(\xi_i - \hat{A}_i \nabla f(x_i))\right\|^2\right], \tag{190}$$

where $\xi_i = \hat{A}_i(\nabla f(x_i) - g_i)$ is the zero mean effective noise that arises from rescaling the stochastic gradient noise. We may write

$$\mathbb{E}\left[\left\|\sum_{i=1}^{t-1}(\xi_i - \hat{A}_i \nabla f(x_i))\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{i=1}^{t-1}\xi_i - \sum_{i=1}^{t-1}\hat{A}_i \nabla f(x_i)\right\|^2\right] \tag{191}$$

$$= \mathbb{E}\left[\left\|\sum_{i=1}^{t-1}\hat{A}_i \nabla f(x_i)\right\|^2 + \left\|\sum_{i=1}^{t-1}\xi_i\right\|^2 - 2\sum_{i=1}^{t-1}\sum_{j=1}^{t-1}\langle \xi_i, \hat{A}_j \nabla f(x_j)\rangle\right] \tag{192}$$

$$= \mathbb{E}\left[\left\|\sum_{i=1}^{t-1}\hat{A}_i \nabla f(x_i)\right\|^2\right] + \mathbb{E}\left[\left\|\sum_{i=1}^{t-1}\xi_i\right\|^2\right] \tag{193}$$

because ξ_i are zero mean. Since $\mathbb{E}[\xi_i^T \xi_j] = 0$ for $i \neq j$, the expression can be simplified as:

$$\mathbb{E}\left[\left\|\sum_{i=1}^{t-1}(\xi_i - \hat{A}_i \nabla f(x_i))\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{i=1}^{t-1}\hat{A}_i \nabla f(x_i)\right\|^2\right] + \sum_{i=1}^{t-1}\mathbb{E}\left[\|\xi_i\|^2\right] \tag{194}$$

$$\leq \mathbb{E}\left[\left\|\sum_{i=1}^{t-1}\hat{A}_i \nabla f(x_i)\right\|^2\right] + \sum_{i=1}^{t-1}\mathbb{E}\left[\|\xi_i\|^2\right] \tag{195}$$

$$\leq \mathbb{E}\left[\left(\sum_{i=1}^{t-1}\left\|\hat{A}_i \nabla f(x_i)\right\|\right)^2\right] + \sum_{i=1}^{t-1}\mathbb{E}\left[\|\xi_i\|^2\right] \tag{196}$$

$$\leq (t-1)\sum_{i=1}^{t-1}\mathbb{E}\left[\left\|\hat{A}_i \nabla f(x_i)\right\|^2\right] + \sum_{i=1}^{t-1}\mathbb{E}\left[\|\xi_i\|^2\right]. \tag{197}$$

Note

$$\mathbb{E}[\|\xi_i\|^2] \leq \mathbb{E}[\|\hat{A}_i \nabla f(x_i)\|^2] + \mathbb{E}[\|\hat{A}_i g_i\|^2] \tag{198}$$

$$\leq \mathbb{E}[\|\hat{A}_i \nabla f(x_i)\|^2] + \frac{9}{4}\Gamma \tag{199}$$

where we have used Lemma E.15. We can then bound

$$\mathbb{E}\left[\left\|\sum_{i=1}^{t-1}(\xi_i - \hat{A}_i \nabla f(x_i))\right\|^2\right] \leq (t-1+1)\sum_{i=1}^{t-1}\mathbb{E}\left[\left\|\hat{A}_i \nabla f(x_i)\right\|^2\right] + \frac{9t\Gamma}{4}. \tag{200}$$

Plugging in Equation (188) we get:

$$\mathbb{E}\left[\left\|\sum_{i=1}^{t-1}(\xi_i - \hat{A}_i \nabla f(x_i))\right\|^2\right] \leq t\left(\frac{\eta L \Gamma \Lambda_1 t}{2} + \frac{2 f_{\text{thresh}}\Lambda_1}{\eta}\right) + t\Gamma. \tag{201}$$

Plugging this into Equation (190) yields:

$$\mathbb{E}[\|x_t - x_1\|^2] \leq t\eta^2\left(\frac{\eta L \Gamma \Lambda_1 t}{2} + \frac{2 f_{\text{thresh}}\Lambda_1}{\eta}\right) + \eta^2 \Gamma t \tag{202}$$

$$= \left(4\eta f_{\text{thresh}}\Lambda_1 + \eta^2\Gamma\right)t + \frac{\eta^3 L \Gamma \Lambda_1 t^2}{2}. \tag{203}$$

Then we may write

$$\mathbb{E}[\|x_t - x_0\|^2] \leq 2\,\mathbb{E}[\|x_t - x_1\|^2] + 2\,\mathbb{E}[\|x_1 - x_0\|^2] \tag{204}$$

$$\leq \left(4\eta f_{\text{thresh}}\Lambda_1 + 2\eta^2\Gamma\right)t + \eta^3 L \Gamma \Lambda_1 t^2 + 2\Gamma r^2. \tag{205}$$

We are almost done. By our additional assumption that $\eta \leq \frac{f_{\text{thresh}}\Lambda_1}{\Gamma}$ (which will wind up being true for small enough τ), it also follows that

$$2\eta^2\Gamma \leq 2\eta f_{\text{thresh}}\Lambda_1 \tag{206}$$

and therefore

$$\mathbb{E}[\|x_t - x_0\|^2] \leq 6\eta f_{\text{thresh}}\Lambda_1 t + \eta^3 L \Gamma \Lambda_1 t^2 + 2\Gamma r^2. \tag{207}$$

□

E.6. Auxiliary lemmas

Lemma E.12. *For $z, A, B, C \geq 0$,*

$$\sqrt{Az^2 + Bz + C} \leq \sqrt{A} \cdot \left(2z + \frac{B}{2A} + \sqrt{\frac{C}{A}}\right). \tag{208}$$

Proof. Note the following two facts:

$$Az^2 + Bz + C = A(z^2 + B/Az + C/A) = A[(z + B/(2A))^2 + C/A - B^2/(2A)^2] \tag{209}$$

and

$$Az^2 + Bz + C = A(z^2 + B/Az + C/A) = A[(z + \sqrt{C/A})^2 - 2\sqrt{C/A} + B/A]. \tag{210}$$

If $B^2 \geq 4AC$, then $C/A - B^2/(2A)^2 \leq 0$. Otherwise, $-2\sqrt{C/A} + B/A \leq 0$. Hence,

$$\sqrt{Az^2 + Bz + C} \leq \begin{cases} \sqrt{A} \cdot (z + B/(2A)) & \text{case 1} \\ \sqrt{A} \cdot (z + \sqrt{C/A}) & \text{case 2.} \end{cases} \tag{211}$$

$$\leq \sqrt{A} \cdot \left[(z + B/(2A)) + (z + \sqrt{C/A})\right]. \tag{212}$$

□

Lemma E.13. *Let $0 < x < 1$. For $t \geq 2 \log C/x$, we have $(1+x)^t \geq C$.*

Proof. For $x < 1$ we have $\log(1+x) \leq x - x^2/2 \leq x/2$. Hence,

$$t \log(1+x) \geq tx/2 \tag{213}$$

$$\geq \log C, \tag{214}$$

and the lemma follows by exponentiating both sides. □

E.6.1. Series lemmas

Lemma E.14 (As in Daneshmand et al. (2018))**.** *For $0 < \beta < 1$ the following inequalities hold:*

$$\sum_{i=1}^{t} (1+\beta)^{t-i} \leq 2\beta^{-1}(1+\beta)^t \tag{215}$$

$$\sum_{i=1}^{t} (1+\beta)^{t-i} i \leq 2\beta^{-2}(1+\beta)^t \tag{216}$$

$$\sum_{i=1}^{t} (1+\beta)^{t-i} i^2 \leq 6\beta^{-3}(1+\beta)^t. \tag{217}$$

E.7. Descent lemmas

First we need a quick lemma relating the constants of the true preconditioner to those of an approximate preconditioner:

Lemma E.15. *Let Γ be an upper bound on $\mathbb{E}[\|Ag\|^2]$. Let $\hat{A}$ be another matrix with $\|\hat{A} - A\| \leq \Delta < \lambda_-/2$. Then, $\mathbb{E}[\|\hat{A}g\|^2] \leq \frac{9}{4}\Gamma$.*

Proof. The proof is straightforward:

$$\mathbb{E}[\|\hat{A}g\|^2] \leq \mathbb{E}[\|(A + \Delta I)g\|^2] \tag{218}$$

$$\leq \mathbb{E}\left[\left\|\frac{3}{2}Ag\right\|^2\right] \tag{219}$$

$$= \frac{9}{4}\,\mathbb{E}[\|Ag\|^2] = \frac{9}{4}\Gamma \tag{220}$$

where the penultimate line follows by $\Delta < \lambda_-/2$ and $\Delta I \preceq \frac{1}{2}A_t$. □

Note that in the noiseless case $\Delta = 0$, all the below results still apply, and we only lose a constant factor compared to the typical descent lemma.

Lemma E.16. *Assume f has L-Lipschitz gradient. Suppose we perform the updates $x_{t+1} \leftarrow x_t - \eta \hat{A}_t g_t$, where g_t is a stochastic gradient, A_t is a $(\Lambda_1, \Lambda_2, \Gamma, \nu, \lambda_-)$-preconditioner, and $\|\hat{A}_t - A_t\| \le \Delta < \frac{\lambda_-}{2}$. Then,*

$$\mathbb{E}[f(x_{t+1})] \le f(x_t) - \frac{\eta\lambda_-}{2}\|\nabla f(x_t)\|^2 + \frac{9\eta^2 L\Gamma}{8} \tag{221}$$

Proof. We write

$$\begin{aligned}\mathbb{E}[f(x_{t+1})] &\le f(x_t) + \langle \nabla f(x_t), \mathbb{E}[x_{t+1} - x_t]\rangle + \frac{L}{2}\mathbb{E}[\|x_{t+1} - x_t\|^2] && (222)\\ &= f(x_t) - \eta\langle \nabla f(x_t), \hat{A}_t \nabla f(x_t)\rangle + \frac{\eta^2 L}{2}\mathbb{E}\left[\|\hat{A}_t g_t\|^2\right] && (223)\\ &\le f(x_t) - \eta(\lambda_- - \Delta)\|\nabla f(x_t)\|^2 + \frac{9\eta^2 L\Gamma}{8} && (224)\\ &\le f(x_t) - \frac{\eta\lambda_-}{2}\|\nabla f(x_t)\|^2 + \frac{9\eta^2 L\Gamma}{8} && (225)\end{aligned}$$

where the third line follows by Lemma E.15. □

Corollary E.2. *Always*

$$\mathbb{E}[f(x_1)] - f(x_0) \le \frac{9\eta^2 L\Gamma}{8}. \tag{226}$$

Corollary E.3. *Suppose $\eta \le 4\lambda_-\|\nabla f(x_0)\|^2/(9L\Gamma)$. Then,*

$$\mathbb{E}[f(x_1)] - f(x_0) \le -\frac{\eta\lambda_-}{4}\|\nabla f(x_0)\|^2. \tag{227}$$

Corollary E.4. *Suppose $\|\nabla f(x_0)\|^2 \ge \tau^2$. Then if $\eta \le 4\lambda_-\tau^2/(9L\Gamma)$*

$$\mathbb{E}[f(x_1)] - f(x_0) \le -\frac{\eta\lambda_-}{4}\|\nabla f(x_0)\|^2 \le -\frac{\eta\lambda_-}{4}\tau^2. \tag{228}$$

F. Convergence to First-Order Stationary Points

F.1. Generic Preconditioners: Proof of Theorem 4.2

Proof. Let g be the stochastic gradient at time t. We will precondition by $A_t = A(x_t)$. We write

$$\begin{aligned}\mathbb{E}[f(x_{t+1})] &\le f(x_t) + \langle \nabla f(x_t), \mathbb{E}[x_{t+1} - x_t]\rangle + \frac{L}{2}\mathbb{E}[\|x_{t+1} - x_t\|^2] && (229)\\ &= f(x_t) - \eta\langle \nabla f(x_t), A_t \nabla f(x_t)\rangle + \frac{\eta^2 L}{2}\mathbb{E}\left[\|A_t g_t\|^2\right] && (230)\\ &\le f(x_t) - \eta\langle \nabla f(x_t), A_t \nabla f(x_t)\rangle + \frac{\eta^2 L\Gamma}{2} && (231)\\ &\le f(x_t) - \eta\lambda_{\min}(A_t)\|\nabla f(x_t)\|^2 + \frac{\eta^2 L\Gamma}{2} && (232)\\ &\le f(x_t) - \eta\lambda_-\|\nabla f(x_t)\|^2 + \frac{\eta^2 L\Gamma}{2}. && (233)\end{aligned}$$

Summing and telescoping, we have

$$\mathbb{E}[f(x_T)] \le \mathbb{E}[f(x_0)] - \eta\lambda_- \sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + \frac{\eta^2 LT\Gamma}{2}. \tag{234}$$

Now rearrange, and bound $f(x_T)$ by f^* to get:

$$\frac{1}{T} \cdot \lambda_- \cdot \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \frac{f(x_0) - f^*}{T\eta} + \frac{\eta L \Gamma}{2}. \tag{235}$$

and therefore

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \left(\frac{f(x_0) - f^*}{T\eta} + \frac{\eta L \Gamma}{2}\right) \cdot \frac{1}{\lambda_-}. \tag{236}$$

Optimally choosing $\eta = \sqrt{2(f(x_0) - f^*)/(TL\Gamma)}$ yields the overall bound

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \sqrt{\frac{2(f(x_0) - f^*)L\Gamma}{T}} \cdot \frac{1}{\lambda_-}. \tag{237}$$

Rephrasing, in order to be guaranteed that the left hand term is bounded by τ^2, it suffices to choose T so that

$$\sqrt{\frac{2(f(x_0) - f^*)L\Gamma}{T}} \cdot \frac{1}{\lambda_-} \leq \tau^2 \tag{238}$$

$$\Leftrightarrow T \geq \frac{2(f(x_0) - f^*)L\Gamma}{\tau^4 \lambda_-^2} \tag{239}$$

and

$$\eta = \sqrt{\frac{2(f(x_0) - f^*)}{TL\Gamma}} \tag{240}$$

$$\leq \sqrt{\frac{2(f(x_0) - f^*)}{L\Gamma} \cdot \frac{\tau^4 \lambda_-^2}{2(f(x_0) - f^*)L\Gamma}} = \frac{\tau^2 \lambda_-}{L\Gamma}. \tag{241}$$

□

F.2. Generic Preconditioners with Errors: Proof of Theorem 4.3

Proof. Let g be the stochastic gradient at time t. We will precondition by $\hat{A}_t$ which satisfies $\|\hat{A}_t - A_t\| \leq \Delta < \lambda_-/2$. We write

$$\mathbb{E}[f(x_{t+1})] \leq f(x_t) + \langle \nabla f(x_t), \mathbb{E}[x_{t+1} - x_t] \rangle + \frac{L}{2} \mathbb{E}[\|x_{t+1} - x_t\|^2] \tag{242}$$

$$= f(x_t) - \eta \langle \nabla f(x_t), \hat{A}_t \nabla f(x_t) \rangle + \frac{\eta^2 L}{2} \mathbb{E}\left[\|\hat{A}_t g_t\|^2\right] \tag{243}$$

$$\leq f(x_t) - \eta(\lambda_- - \Delta)\|\nabla f(x_t)\|^2 + \frac{\eta^2 L}{2} \mathbb{E}\left[\|(A_t + \Delta I)g_t\|^2\right] \tag{244}$$

$$\leq f(x_t) - \frac{\eta \lambda_-}{2}\|\nabla f(x_t)\|^2 + \frac{\eta^2 L}{2} \mathbb{E}\left[\left\|\frac{3}{2} A_t g_t\right\|^2\right] \tag{245}$$

$$= f(x_t) - \frac{\eta \lambda_-}{2}\|\nabla f(x_t)\|^2 + \frac{9\eta^2 L \Gamma}{8} \tag{246}$$

where the penultimate line follows by $\Delta < \lambda_-/2$ and $\Delta I \preceq \frac{1}{2} A_t$. Summing and telescoping, and further bounding $9/8 < 2$, we have

$$\mathbb{E}[f(x_T)] \leq \mathbb{E}[f(x_0)] - \frac{\eta \lambda_-}{2} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] + 2\eta^2 L \Gamma. \tag{247}$$

Now rearrange, and bound $f(x_T)$ by f^* to get:

$$\frac{1}{T} \cdot \frac{\lambda_-}{2} \cdot \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \frac{f(x_0) - f^*}{T\eta} + 2\eta L\Gamma \tag{248}$$

and therefore

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \left(\frac{f(x_0) - f^*}{T\eta} + 2\eta L\Gamma\right) \frac{2}{\lambda_-}. \tag{249}$$

Optimally choosing $\eta = \sqrt{(f(x_0) - f^*)/(2TL\Gamma)}$ yields the overall bound

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \sqrt{\frac{32(f(x_0) - f^*)L\Gamma}{T}} \cdot \frac{1}{\lambda_-}. \tag{250}$$

Rephrasing, in order to be guaranteed that the left hand term is bounded by τ^2, it suffices to choose T so that

$$\sqrt{\frac{32(f(x_0) - f^*)L\Gamma}{T}} \cdot \frac{1}{\lambda_-} \leq \tau^2 \tag{251}$$

$$\Leftrightarrow T \geq \frac{32(f(x_0) - f^*)L\Gamma}{\tau^4 \lambda_-^2} \tag{252}$$

and

$$\eta = \sqrt{\frac{f(x_0) - f^*}{2TL\Gamma}} \leq \sqrt{\frac{(f(x_0) - f^*)\tau^4\lambda_-^2}{32(f(x_0) - f^*)L^2\Gamma^2}} = \frac{\tau^2\lambda_-}{4\sqrt{2}L\Gamma}. \tag{253}$$

□

G. Online Matrix Estimation

We first reproduce the Matrix Bernstein inequality as presented by Tropp (2015):

Theorem G.1 (Matrix Bernstein)**.** *Let $Z_i \in \mathbb{R}^{d\times d}$ for $i = 1, \ldots, n$ be independent, random, symmetric matrices. Assume $\mathbb{E}[Z_i] = 0$ and $\|Z_i\| \leq R$ for all i. Define $v = \|\sum_{i=1}^n \mathbb{E}[Z_i^2]\|$. Then for all $k \geq 0$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n Z_i\right\| \geq k\right) \leq d \exp\left(\frac{-k^2/2}{v + Rk/3}\right).$$

Corollary G.1. *Let $Z_i \in \mathbb{R}^{d\times d}$ for $i = 1, \ldots, n$ be independent, random, symmetric matrices. Assume $\mathbb{E}[Z_i] = 0$ and $\|Z_i\| \leq R$ for all i. Assume $\|\mathbb{E}[Z_i^2]\| \leq \sigma_{\max}^2$. Let $w \in \Delta_n$ in the simplex. Then for all $k \leq 3\|w\|_2^2\sigma_{\max}^2/R$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n w_i Z_i\right\| \geq k\right) \leq d \exp\left(\frac{-k^2}{4\|w\|_2^2\sigma_{\max}^2}\right).$$

Proof. Simply apply Theorem G.1 with $\hat{Z}_i = w_i Z_i$. □

Now we can apply the above matrix concentration results to prove Theorem 4.1:

Proof of Theorem 4.1. First we separately bound the bias and variance, then use Corollary G.1. The bias is:

$$\left\|\sum_{t=1}^{T} w_t f(x_t) - f(x_T)\right\| = \left\|\sum_{t=1}^{T} w_t (f(x_t) - f(x_T))\right\| \tag{254}$$

$$\leq \sum_{t=1}^{T} w_t \|f(x_t) - f(x_T)\| \tag{255}$$

$$\leq L \sum_{t=1}^{T} w_t \|x_t - x_T\| \tag{256}$$

$$\leq L \sum_{t=1}^{T} w_t \sum_{s=t+1}^{T} \|x_s - x_{s-1}\| \tag{257}$$

$$\leq \eta M L \sum_{t=1}^{T} w_t (T-t) \tag{258}$$

$$= \eta M L \cdot \frac{1}{\sum_{t=1}^{T} \beta^{T-t}} \cdot \sum_{t=1}^{T} \beta^{T-t}(T-t). \tag{259}$$

Note that by a well-known identity,

$$\sum_{t=1}^{T} \beta^{T-t}(T-t) = \sum_{s=0}^{T-1} s\beta^s \leq \sum_{s=0}^{\infty} s\beta^s = \frac{\beta}{(1-\beta)^2}. \tag{260}$$

Hence, the bias is bounded by

$$\eta M L \cdot \frac{1}{\sum_{t=1}^{T} \beta^{T-t}} \cdot \frac{\beta}{(1-\beta)^2} = \eta M L \cdot \frac{1-\beta}{1-\beta^T} \cdot \frac{\beta}{(1-\beta)^2} \tag{261}$$

$$= \eta M L \cdot \frac{1}{1-\beta^T} \cdot \frac{\beta}{1-\beta} \tag{262}$$

$$\leq M L \cdot \frac{\eta}{(1-\beta)(1-\beta^T)}. \tag{263}$$

Applying Corollary G.1 to $Z_t = Y_t - f(x_t)$, we have that

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T} w_t (Y_t - f(x_t))\right\| > k\right) \leq d \exp\left(\frac{-k^2}{4\|w\|_2^2 \sigma_{\max}^2}\right).$$

Now note that

$$\|w\|_2^2 = \sum_{t=1}^{T} w_t^2 = \frac{1}{(\sum_{t=1}^{T} \beta^{T-t})^2} \sum_{t=1}^{T} (\beta^2)^{T-t} \tag{264}$$

$$= \frac{(1-\beta)^2}{(1-\beta^T)^2} \sum_{t=1}^{T} (\beta^2)^{T-t} \tag{265}$$

$$= \frac{(1-\beta)^2}{(1-\beta^T)^2} \cdot \frac{1-\beta^{2T}}{1-\beta^2} \tag{266}$$

$$= \frac{1-\beta^{2T}}{(1-\beta^T)^2} \cdot \frac{(1-\beta)^2}{1-\beta^2} \tag{267}$$

$$= \frac{1+\beta^T}{1-\beta^T} \cdot \frac{1-\beta}{1+\beta} \tag{268}$$

$$\leq \frac{2(1-\beta)}{1-\beta^T}. \tag{269}$$

Setting the right hand side of the high probability bound to δ, we have concentration w.p. $1-\delta$ for k satisfying

$$\delta \geq d \exp\left(\frac{-k^2}{4\|w\|_2^2 \sigma_{\max}^2}\right). \tag{270}$$

Rearranging, we find

$$\log(d/\delta) \leq \frac{k^2}{4\|w\|_2^2 \sigma_{\max}^2} \tag{271}$$

$$\Leftrightarrow k \geq 2\sigma_{\max}\|w\|_2 \sqrt{\log(d/\delta)}. \tag{272}$$

Combining this with the triangle inequality,

$$\left\|\sum_{t=1}^{T} w_t y_t - f(x_T)\right\| \leq \left\|\sum_{t=1}^{T} w_t y_t - \sum_{t=1}^{T} w_t f(x_t)\right\| + \left\|\sum_{t=1}^{T} w_t (f(x_t) - f(x_T))\right\| \tag{273}$$

$$\leq 2\sigma_{\max}\|w\|_2 \sqrt{\log(d/\delta)} + ML \cdot \frac{\eta}{(1-\beta)(1-\beta^T)} \tag{274}$$

$$\leq 2^{3/2}\sigma_{\max} \frac{\sqrt{1-\beta}}{\sqrt{1-\beta^T}} \sqrt{\log(d/\delta)} + ML \cdot \frac{\eta}{(1-\beta)(1-\beta^T)}. \tag{275}$$

with probability $1-\delta$. Since $1/\sqrt{1-\beta^T} \leq 1/(1-\beta^T)$, this can further be bounded by

$$\left(2^{3/2}\sigma_{\max}\sqrt{1-\beta}\sqrt{\log(d/\delta)} + ML \cdot \frac{\eta}{(1-\beta)}\right) \cdot \frac{1}{1-\beta^T}. \tag{276}$$

Write $\alpha = 1-\beta$. The inner part of the bound is optimized when

$$2^{3/2}\sigma_{\max}\sqrt{\alpha}\sqrt{\log(d/\delta)} = ML \cdot \frac{\eta}{\alpha} \tag{277}$$

$$\Leftrightarrow \alpha^{3/2} = \frac{ML\eta}{2^{3/2}\sigma_{\max}\sqrt{\log(d/\delta)}} \tag{278}$$

$$\Leftrightarrow \alpha = \frac{M^{2/3}L^{2/3}\eta^{2/3}}{2\sigma_{\max}^{2/3}(\log(d/\delta))^{1/3}} \tag{279}$$

for which the overall inner bound is

$$2 \cdot 2^{3/2}\sigma_{\max}\sqrt{\alpha}\sqrt{\log(d/\delta)} = 4\sigma_{\max}^{2/3}(\log(d/\delta))^{1/3} M^{1/3} L^{1/3} \eta^{1/3}. \tag{280}$$

If T is sufficiently large, the $1/(1-\beta^T)$ term will be less than 2. In particular,

$$T > \frac{2}{\log(1+\alpha)} \implies \frac{1}{1-(1-\alpha)^T} < 2. \tag{281}$$

Since $\log(1+\alpha) > \alpha/2$ for $\alpha < 1$, it suffices to have $T > 4/\alpha$. □

H. Converting Noise Estimates into Preconditioner Estimates

Lemma H.1. *Suppose $\|G - \hat{G}\| \leq \varepsilon$, i.e. $\hat{G}$ is a good estimate of G in operator norm. Assume ε is so small that $\varepsilon\|G^{-1}\| < 1/2$. Then,*

$$\|G^{-1} - \hat{G}^{-1}\| \leq \frac{\varepsilon}{2(\lambda_{\min}(G))^2}. \tag{282}$$

Proof. Observe

$$G^{-1}(\hat{G} - G)\hat{G}^{-1} = G^{-1} - \hat{G}^{-1}. \tag{283}$$

Therefore,

$$\delta = \|G^{-1} - \hat{G}^{-1}\| = \|G^{-1}(\hat{G} - G)\hat{G}^{-1}\| \tag{284}$$
$$\leq \varepsilon\|G^{-1}\|\|\hat{G}^{-1}\| \tag{285}$$
$$\leq \varepsilon\|G^{-1}\|(\|G^{-1}\| + \delta). \tag{286}$$

Grouping δ terms together, we find

$$(1 - \varepsilon\|G^{-1}\|)\delta \leq \varepsilon\|G^{-1}\|^2 \tag{287}$$
$$\implies \delta \leq \frac{\|G^{-1}\|^2}{1 - \varepsilon\|G^{-1}\|} \cdot \varepsilon. \tag{288}$$

By assumption ε is small enough so that $\varepsilon\|G^{-1}\| < 1/2$, so overall we have

$$\delta \leq \frac{\|G^{-1}\|^2}{2} \cdot \varepsilon = \frac{1}{2(\lambda_{\min}(G))^2} \cdot \varepsilon. \tag{289}$$

□

Lemma H.2. *Suppose $\|G - \hat{G}\| \leq \varepsilon$, i.e. $\hat{G}$ is a good estimate of G in operator norm. Assume ε is so small that $\varepsilon < \frac{3}{4}\lambda_{\min}(G)$. Then,*

$$\|G^{1/2} - \hat{G}^{1/2}\| \leq \frac{\varepsilon}{(\lambda_{\min}(G))^{1/2}}. \tag{290}$$

Proof. We can equivalently write

$$G - \varepsilon I \preceq \hat{G} \preceq G + \varepsilon I. \tag{291}$$

By monotonicity of the matrix square root,

$$(G - \varepsilon I)^{1/2} \preceq \hat{G}^{1/2} \preceq (G + \varepsilon I)^{1/2} \tag{292}$$

and therefore

$$(G - \varepsilon I)^{1/2} - G^{1/2} \preceq \hat{G}^{1/2} - G^{1/2} \tag{293}$$
$$\preceq (G + \varepsilon I)^{1/2} - G^{1/2}. \tag{294}$$

At this point we can bound each side by applying Lemma H.3 to G and to $G - \varepsilon I$. The result is the bound

$$\frac{-\varepsilon}{2(\lambda_{\min}(G) - \varepsilon)^{1/2}} \preceq \hat{G}^{1/2} - G^{1/2} \preceq \frac{\varepsilon}{2(\lambda_{\min}(G))^{1/2}}.$$

The lower bound is looser, so the operator norm of the difference is bounded by

$$\frac{\varepsilon}{2(\lambda_{\min}(G) - \varepsilon)^{1/2}} < \frac{\varepsilon}{2(\frac{1}{4}\lambda_{\min}(G))^{1/2}} = \frac{\varepsilon}{(\lambda_{\min}(G))^{1/2}}.$$

□

Lemma H.3. *Let $A \succ 0$ and $\varepsilon > 0$. Then*

$$\|(A + \varepsilon)^{1/2} - A^{1/2}\| \leq \frac{\varepsilon}{2(\lambda_{\min}(A))^{1/2}}. \tag{295}$$

Proof. The bound reduces to plugging in the eigenvalues of A to a scalar function $f : \mathbb{R} \to \mathbb{R}$. Define $f(x) = (x+\varepsilon)^{1/2} - x^{1/2}$. Note that

$$f(x) = \frac{((x+\varepsilon)^{1/2} - x^{1/2})((x+\varepsilon)^{1/2} + x^{1/2})}{(x+\varepsilon)^{1/2} + x^{1/2}} \tag{296}$$

$$= \frac{(x+\varepsilon) - x}{(x+\varepsilon)^{1/2} + x^{1/2}} \tag{297}$$

$$= \frac{\varepsilon}{(x+\varepsilon)^{1/2} + x^{1/2}} \tag{298}$$

$$\leq \frac{\varepsilon}{2x^{1/2}}, \tag{299}$$

from which the result follows. □

Corollary H.1. *Suppose $\|G - \hat{G}\| \leq \varepsilon$, for small enough ε. Then,*

$$\|(G+\delta I)^{-1/2} - (\hat{G}+\delta I)^{-1/2}\| \leq \frac{\varepsilon}{2(\delta + \lambda_{\min}(G))^{3/2}}.$$

Proof. Simply apply Lemma H.1 and Lemma H.2 to $G + \delta I$. □