Acceleration with Stochastic Linear Coupling

Benjamin Paul-Dubois-Taine
Student Number 70300892
Department of Computer Science
University of British Columbia

Victor S. Portella
Student Number 93784759
Department of Computer Science
University of British Columbia

Sikander Randhawa
Student Number 35808120
Department of Computer Science
University of British Columbia

Abstract

We prove that the linear coupling algorithm of Allen-Zhu and Orecchia [2014] preserves its accelerated convergence rate even if it only has oracle access to noisy gradients satisfying a strong growth condition. Acceleration has already been attained with such an oracle, however our analysis yields a slightly more transparent proof by avoiding the standard analysis of Nesterov's accelerated method.

1 Introduction

Most modern machine learning algorithms are trained using *stochastic* first-order optimization methods. This is because deterministic methods, like gradient descent (GD), have prohibitively large iteration costs as each iteration requires full gradient computation – an issue when working with huge datasets. Theoretically, stochastic methods require many more iterations to converge to an optimal solution than their deterministic counterparts. This may potentially offset the benefit enjoyed by stochastic methods of reduced computational complexity per iteration. Empirically however, stochastic gradient descent (SGD) seems to enjoy the same iteration complexity as GD on a large variety of tasks. This has led researchers to try and exploit the structure of modern machine learning models to explain this behavior. One key observation is that with the increase of computational capability in the past few years, many machine learning models used in practice are able to *interpolate* the data, namely fit the data perfectly. This is true in particular for deep neural networks as shown by Zhang et al. [2016]. Vaswani et al. [2018] show that finite sum models that interpolate the data satisfy a *weak growth* condition (WGC), and in the same work that SGD obtains rates matching that of GD under weak growth. This result offers an explanation of the good performance of SGD in practice.

With stochastic rates matching the deterministic rates under this weak growth condition, it is natural to ask if there exists stochastic algorithms whose rates match the rates of Nesterov's accelerated GD in the presence of interpolation. Although acceleration (without variance-reduction) when conditions on the gradient noise are guaranteed has been thoroughly studied (see the work by Cohen et al. [2018] and references therein), this remains an open question when only interpolation/weak growth is guaranteed. However, Vaswani et al. [2018] laid the first stone towards solving this by proving that Nesterov's accelerated method retains its optimal convergence rate even if it uses noisy gradients satisfying the *strong growth* condition (SGC) – a stronger condition than weak growth which was previously studied by Schmidt and Roux [2013]. The proof of this result mimics the analysis from the deterministic setting, which is known to be notoriously opaque.

Our work is primarily motivated by the ultimate goal of obtaining accelerated rates under the practical assumption of weak growth. Such a result would indicate that the weak growth condition captures

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

the essence of why stochastic methods perform surprisingly well when applied to many machine learning problems. We analyse a stochastic variant of the linear coupling algorithm proposed by Allen-Zhu and Orecchia [2014]. The analysis of linear coupling essentially provides a framework for combining the analyses of algorithms which satisfy two key properties. (We will elaborate on this in Subsection 2.3 where we outline the high level analysis of our main result.) It is conceivable that this analytic framework is more amenable to proving accelerated convergence rates in the stochastic setting, especially considering the comparatively challenging analysis of Nesterov's method which was the previous approach to obtaining acceleration under SGC.

Main result. This is indeed true under the strong growth assumption. We show that a stochastic variant of the linear coupling algorithm enjoys an accelerated rate of convergence. Although acceleration under strong growth is not a new result, our hope is that this framework will also prove to be useful under WGC (for more on this, see Section 4) since it offers a fairly simple analysis.

The paper is organized as follows. In Subsection 1.1 we present some related work in the field of stochastic first-order methods in machine learning. In Section 2 we lay out the assumptions underlying our analysis, present the algorithm, and outline the approach of our analysis. In Section 3 we prove a slightly more general version of our main result. Finally we conclude by a brief discussion of ideas for future work.

1.1 Related Work

When optimizing a smooth convex function, it is well-known that the error of classical GD is $O(1/k)$ after k iterations. However, such a convergence rate is not optimal for first-order methods as shown by Nesterov [2004]. A modified version of gradient descent proposed by Nesterov [1983], known as Nesterov's accelerated GD, achieves the optimal convergence rate of $O(1/k^2)$. One drawback of this algorithm is it involves an intricate analysis and lacks intuition. This makes it challenging to modify and extend to different settings. In recent years there has been a surge of research that propose simplified analyses of Nesterov's acceleration or define new accelerated methods (i.e. with a $O(1/k^2)$ rate) in order to illuminate the key ideas of acceleration [Allen-Zhu and Orecchia, 2014, Su et al., 2014, Bubeck et al., 2015, Krichene et al., 2015, Diakonikolas and Orecchia, 2019]. Our work extends this latter line of research by augmenting the *linear coupling* algorithm due to Allen-Zhu and Orecchia [2014] to propose a clearer accelerated method in the stochastic setting.

In the stochastic setting (where the method uses random estimates of the gradient), SGD achieves only a (expected) convergence rate of $O(1/\sqrt{k})$ for smooth and convex functions [Nesterov, 2004], which is far worse than the theoretical guarantees of its non stochastic counterpart. A thoroughly studied special case is the *finite-sum* setting where the objective function is the average of (finitely-many) smooth convex functions and the stochastic gradient at each iteration is the gradient of one of these functions chosen uniformly at random. In such a case, one can attain a $O(1/k)$ convergence rate with SGD with a technique known as *variance-reduction* [Johnson and Zhang, 2013, Shalev-Shwartz and Zhang, 2013, Schmidt et al., 2017]. By using a variance-reduced method together with the meta-algorithm from Lin et al. [2015] one can already achieve a convergence rate of $O((\log k)/k^2)$, only a $O(\log k)$ factor away from the optimal accelerated rate. Allen-Zhu [2017] showed how to achieve a $O(1/k^2)$ convergence rate by combining variance-reduction ideas with the linear coupling algorithm. However, these methods is that they need to sporadically compute deterministic gradients, and/or require the finite-sum structure assumption.

More closely related to this work is the study of stochastic first-order methods when the model being optimized can *interpolate*[1] the data [Schmidt and Roux, 2013, Vaswani et al., 2018, Ma et al., 2018]. In particular, Vaswani et al. [2018] show that, under a strong notion of interpolation known as the strong growth condition, Nesterov's accelerated GD achieves a $O(1/k^2)$ convergence rate. Our work is closely related to this latter result, since we propose a stochastic version of linear coupling, which possesses a simpler analysis than Nesterov's accelerated GD and achieves the same accelerated rate under identical conditions.

[1] In the finite-sum case, this means that the global minimum of the objective function in the finite sum setting is also a global minimum of each individual component function. This idea can be formally extended to non-finite-sum objective functions in different ways as Vaswani et al. [2018] show.

2 Algorithm and Results

In this section, we describe the set of assumptions we will be working with, followed by a formal statement of our result. We will then provide a high-level description of the main ideas of our analysis.

2.1 Preliminaries

Notation and standard assumptions. Let $f : \mathbb{R}^d \to R$ be a convex and differentiable function. We will assume that f is smooth, i.e. ∇f is L-Lipschitz continuous for some constant $L > 0$. Further, we assume that f reaches a minimum at x^*. Unless stated otherwise, we will use $\|\cdot\|$ to refer to the Euclidean norm on $\mathbb{R}^d$ (i.e. $\|x\| = \sqrt{\langle x, x\rangle}$).

Bregman divergence. The linear coupling algorithm makes use of "mirror steps". These require the choice of a regularizer $w \colon \mathbb{R}^d \to \mathbb{R}$ that is differentiable and 1-strongly convex with respect to $\|\cdot\|$ (i.e. $f - \frac{1}{2}\|\cdot\|^2$ is convex). The Bregman divergence (w.r.t. w) is then defined as

$$D_w(y, x) = w(y) - w(x) - \langle \nabla w(x), y - x\rangle.$$

Noisy gradient oracle. We assume access to a noisy, unbiased, gradient oracle. Formally, for any $x \in \mathbb{R}^d$ and any random vector $u \in \mathbb{R}^d$, we have access to an unbiased estimate $\nabla f(x, u)$ of $\nabla f(x)$, that is, $\mathbb{E}_u\big[\nabla f(x, u)\big] = \nabla f(x)$. The expectation is taken with respect to the distribution from which the vector u is sampled from.

The strong growth condition. The previous assumptions are standard for studying first-order stochastic methods. We now make a crucial assumption on the norm of the stochastic gradients, known as the strong growth condition (SGC), which was previously studied by Schmidt and Roux [2013]. The function f is said to satisfy the SGC with constant ρ if for any $x \in \mathbb{R}^n$ we have

$$\mathbb{E}_u\big[\,\|\nabla f(x, u)\|^2\,\big] \le \rho\,\|\nabla f(x)\|^2\,. \tag{2.1}$$

We emphasize that this assumption is strong and not often satisfied in practice. In particular, in the finite sum setting, SGC reads

$$\mathbb{E}_i\big[\,\|\nabla f_i(x)\|^2\,\big] \le \rho\,\|\nabla f(x)\|^2\,.$$

This implies that if for some $x^* \in \mathbb{R}^d$, $\nabla f(x^*) = 0$, then also $\nabla f_i(x^*) = 0$ for all i. In the case where each component f_i is convex, it implies that the global minimizer of f is also a global minimizer of each individual component function f_i, so that the model is expressive enough to perfectly fit, or *interpolate*, the data.

2.2 Stochastic Linear Coupling

In Algorithm 1 we formally describe the stochastic linear coupling algorithm. Each iteration of Algorithm 1 can be seen as combining an SGD step (line 7) with a stochastic mirror descent (SMD) step (line 8). The statement of the algorithm is almost identical to the deterministic linear coupling algorithm of Allen-Zhu and Orecchia [2014], except for two differences. First of all, it uses noisy gradients instead of full gradients. Moreover, the gradient descent step is taken with respect to the euclidean norm, whereas the linear coupling algorithm as stated by Allen-Zhu and Orecchia [2014] can handle arbitrary norms. Some discussion about the extension of our results to arbitrary norms can be found in Section 4.

Finally, one can show that if the regularizer w is chosen to be $w(x) = \frac{1}{2}\|x\|^2$, Algorithm 1 is identical to the accelerated SGD from Vaswani et al. [2018], as line 8 then reduces to

$$z_{k+1} \leftarrow z_k - \alpha_{k+1}\nabla f(x_{k+1}, u_{k+1})$$

In that sense our approach provides a more general way of finding stochastic first-order algorithms that yields accelerated rates under SGC.

Algorithm 1 Stochastic Linear Coupling

Input: A initial point $x_0 \in \mathbb{R}^n$, a gradient step size $\eta > 0$, a balancing constant $C > 0$, and the number of iterations T.

$y_0 \leftarrow x_0$
$z_0 \leftarrow x_0$
$k \leftarrow 0$
for $k \leftarrow 0$ to $T-1$ **do**
    $\alpha_{k+1} \leftarrow \frac{k+2}{2C}$, and $\tau_k \leftarrow \frac{1}{\alpha_{k+1}C} = \frac{2}{k+2}$.
    $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k) z_k$
    $y_{k+1} \leftarrow x_{k+1} - \eta \nabla f(x_{k+1}, u_{k+1})$
    $z_{k+1} \leftarrow \arg\min_{z \in \mathbb{R}^n} \left\{ D_w(z, z_k) + \langle \alpha_{k+1} \nabla f(x_{k+1}, u_{k+1}), z - z_k \rangle \right\}$
end for
return y_T.

Our main result is that Algorithm 1 enjoys an accelerated rate of convergence under strong growth.

Theorem 2.1. Let $T > 0$, and let $y_T \in \mathbb{R}^d$ be as given by Algorithm 1 with initial point $x_0 \in \mathbb{R}^d$, step size $\eta := 1/L\rho$, and balancing constant $C := L\rho^2 = \frac{\rho}{2\eta(1-\eta L\rho/2)}$. Moreover, let $x^* \in \arg\min_{x \in \mathbb{R}^d} f(x)$ and let $\Theta > 0$ be such that $D_w(x^*, x_0) \leq \Theta$. Then,

$$\mathbb{E}[f(y_T)] - f(x^*) \leq \frac{4\Theta L\rho^2}{(T+1)^2}.$$

Remark 2.2. Assuming a weaker form of strong growth yields a similar result as Theorem 2.1. If there are non-negative constants ρ, σ such that $\mathbb{E}_u \left[\|\nabla f(x,u)\|^2 \right] \leq \rho \|\nabla f(x)\|^2 + \sigma$, then

$$\mathbb{E}[f(y_T)] - f(x^*) \leq \frac{4\Theta L\rho^2}{(T+1)^2} + \left((L\rho\eta)^2 + 1 \right) \frac{2\sigma(T+2)}{3L^2\rho^4}.$$

Asymptotically, this matches the rate obtained by Vaswani et al. [2018] under the same assumption.

The proof of Theorem 2.1 is in Section 3. We proceed to highlight the main ideas of our analysis.

2.3 Main idea of the analysis

The linear coupling algorithm provides an analytic framework for combining the analyses of gradient descent and mirror descent. Loosely first order algorithm provided the existence of two things: 1) an iterative first order algorithm which guarantees a "descent property" in terms of $\|\nabla f(x)\|^2$, and 2) an iterative first order algorithm which can relate the per-round *regret*[2] to the $\|\nabla f(x)\|^2$ term from the descent property. It turns out that gradient descent guarantees property 1) whereas mirror descent guarantees property 2).

Our idea is based off of the fact that the analysis of linear coupling is not dependent on its subroutines actually being exactly gradient descent and mirror descent respectively. Therefore, in the stochastic setting, we simply need to find the right set of algorithms and assumptions which will allow us to obtain (in expectation) properties 1) and 2). From there, we can proceed with the linear coupling analysis – with some tweaks to the step sizes. It turns out that under the strong growth assumption, *stochastic* gradient descent and *stochastic* mirror descent have similar guarantees as their deterministic counterparts. These observations drive the rest of the analysis.

3 Analysis

Theorem 2.1 using the weaker form of strong growth stated in Remark 2.2. Formally, we assume that there are non-negative constants ρ, σ such that

$$\mathbb{E}_u \left[\|\nabla f(x,u)\|^2 \right] \leq \rho \|\nabla f(x)\|^2 + \sigma. \tag{3.1}$$

[2]Regret is a notion of performance in online convex optimization which is the standard performance measure for mirror descent. By averaging iterates on can translate regret guarantees to classical convergence rates.

Moreover, throughout this section we denote by $\mathbb{E}_k$ the expectation conditioned on the noise $u_1, \ldots, u_k$ from Algorithm 1. So, $\mathbb{E}_k$ treats $u_1, \ldots, u_k$ as constants and $u_{k+1}, \ldots, u_T$ as random.

As mentioned before, the main insight behind the analysis of linear coupling is the fact that the per-iteration guarantees of gradient and mirror descent are complementary. Linear coupling combines the complementary key steps of the analyses of gradient and mirror descent.

The main step of the analysis for gradient descent shows that the per iteration decrease of the objective value is proportional to the current squared gradient norm. The next lemma shows that this holds (in expectation) for *stochastic* gradient descent under strong growth.

Lemma 3.1 (Descent lemma)**.** Let $k \geq 0$ and let x_{k+1} and y_{k+1} be defined as in Algorithm 1. Then,

$$\mathbb{E}_k[f(y_{k+1}) - f(x_{k+1})] \leq -\eta\Big(1 - \frac{\eta L \rho}{2}\Big)\|\nabla f(x_{k+1})\| + \frac{\eta^2 \sigma L}{2}$$

Proof. Since f is L-smooth and from the definition of y_{k+1} we have

$$\begin{aligned} f(y_{k+1}) &\leq f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1} \rangle + \frac{L}{2}\|x_{k+1} - y_{k+1}\|^2 \\ &= f(x_{k+1}) - \eta \langle \nabla f(x_{k+1}), \nabla f(x_{k+1}, u_{k+1}) \rangle + \frac{L\eta^2}{2}\|\nabla f(x_{k+1}, u_{k+1})\|^2. \end{aligned}$$

Taking expectation (conditioned on $u_1, \ldots, u_k$) and using (3.1) yields the desired inequality. □

The key step in the analysis of mirror descent bounds the additional regret incurred at each iteration by the squared norm of the gradient plus a penalty related to how far the mirror step went (measured by the regularizer w). An analogue of this property is stated below for *stochastic* mirror descent.

Lemma 3.2 (Mirror descent guarantee, see Allen-Zhu and Orecchia [2014, Appendix B.2])**.** Let $k \geq 0$ and let z_k and z_{k+1} be given by line 8 in Algorithm 1. Then for any $x \in \mathbb{R}^d$,

$$\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - x \rangle \leq \frac{\alpha_{k+1}^2}{2}\mathbb{E}_k[\|\nabla f(x_{k+1}, u_{k+1})\|^2] + D_w(x, z_k) - \mathbb{E}_k[D_w(x, z_{k+1})].$$

Lemma 3.1 and Lemma 3.2 provide the analogues of the descent property and the mirror descent regret guarantee in the stochastic setting under strong growth. To show the accelerated convergence rate we will combine both guarantees in a manner similar to the analysis of the deterministic linear coupling method by Allen-Zhu and Orecchia [2014], .

The following lemma, whose proof we defer to Appendix A shows how one can combine Lemma 3.1 and Lemma 3.2 by using a coupling parameter τ_k, so that the expected change in objective value from round to round telescopes in an appropriate manner. This will yield our main result by summing over all iterations.

Lemma 3.3 (Coupling Lemma)**.** The iterates of Algorithm 1 satisfy the following inequality:

$$\begin{aligned} &\alpha_{k+1}^2 C \mathbb{E}_k[f(y_{k+1})] - (\alpha_{k+1}^2 C - \alpha_{k+1}) f(y_k) \\ &\leq \alpha_{k+1} f(x^*) + \mathbb{E}_k[D_w(x^*, z_{k+1}) - D_w(x^*, z_k)] + (CL\eta^2 + 1)\frac{\alpha_{k+1}^2}{2}\sigma. \end{aligned}$$

The proof of Theorem 2.1 essentially sums the inequality given by Lemma 3.3 over all of iterations.

Proof of Theorem 2.1. First, note that

$$\alpha_k^2 C = \frac{(k+2-1)^2}{4C^2}C = \frac{(k+2)^2}{4C^2}C - \frac{k+2}{2C} + \frac{1}{4C} = \alpha_{k+1}^2 C - \alpha_{k+1} + \frac{1}{4C}.$$

Using the above and summing the inequality from Lemma 3.3 for $k \in \{0, \ldots, T-1\}$ and by taking expectations we have

$$\begin{aligned} &\alpha_T^2 C \mathbb{E}[f(y_T)] + \frac{1}{4C}\sum_{k=1}^{T-1}\mathbb{E}[f(y_k)] \\ &\leq \sum_{k=1}^{T}\alpha_k f(x^*) + D_w(x^*, z_0) - \mathbb{E}[D_w(x^*, z_T)] + \frac{\sigma}{2}\Big(CL\eta^2 + 1\Big)\sum_{k=0}^{T-1}\alpha_{k+1}^2. \end{aligned}$$

Now, by the choice of α_k we have $\sum_{k=1}^{T} \alpha_k = \frac{T(T+3)}{4C}$ and $\sum_{k=1}^{T} \alpha_k^2 \leq \frac{(T+2)^3}{3C^2}$. This together with the non-negativity of Bregman divergence and since x^* is a minimizer of f, we have

$$\frac{(T+1)^2}{4C}\mathbb{E}[f(y_T)] \leq \frac{T(T+3)-T+1}{4C} f(x^*) + \Theta + \Big(CL\eta^2 + 1\Big)\frac{\sigma(T+2)^3}{6C^2}.$$

Multiplying the above by $\frac{4C}{(T+1)^2}$ and using that $\frac{T+2}{T+1} \leq 2$ and that $\eta = 1/L\rho$ we have

$$\begin{aligned}\mathbb{E}[f(y_T)] - f(x^*) &\leq \frac{4\Theta C}{(T+1)^2} + \Big(CL\eta^2 + 1\Big)\frac{4\sigma(T+2)}{6C^2} \\ &= \frac{4\Theta L\rho^2}{(T+1)^2} + \Big((L\rho\eta)^2 + 1\Big)\frac{4\sigma(T+2)}{6L^2\rho^4}.\end{aligned}$$

□

4 Future Work

Non-Euclidean stochastic acceleration. One nice feature of the original linear coupling algorithm by Allen-Zhu and Orecchia [2014] is that its analysis follows seamlessly even if f is smooth with respect to an arbitrary norm instead of the euclidean norm. However, the SGD descent lemma (Lemma 3.1) we used heavily relies on the fact that the euclidean norm is induced by the euclidean inner-product (and that its dual norm is itself). Allen-Zhu [2017] studies stochastic acceleration on non-euclidean settings using ideas from the linear coupling algorithm. However, the techniques used rely on bounding the variance of the noise with variance reduction techniques, and we were not able to extend these results to the strong growth case. An interesting direction of future research would be to extend stochastic linear coupling to the non-Euclidean setting, or even to study the effectiveness of different accelerated stochastic first-order methods such as the one by Cohen et al. [2018] when used with functions that satisfy strong (or weak) growth conditions.

Constrained stochastic optimization.

Acceleration under WGC. One may attempt to use SGD and SMD in the linear coupling framework again under the weaker WGC assumption in an attempt to recover the accelerated rate. An issue here is that deriving a "descent lemma" is not as straightforward under WGC. Moreover, even if a descent property were somehow attainable under WGC, it is not clear how to relate the upper bound from Lemma 3.2 to the $\|\nabla f(x)\|^2$ term from the descent property without the use of strong-growth. We did not spend too much time in this direction – it is possible these ideas are fruitful.

Acceleration under SGC for strongly-convex functions. Our results hold when f is smooth and convex. Another common assumption in the optimization literature is strong-convexity. According to Allen-Zhu and Orecchia [2014], an analysis of their linear coupling algorithm can be made to attain the optimal linear rate in the strongly-convex setting by replacing the standard mirror descent guarantee (Lemma 3.2) with the corresponding analysis for regret minimization of strongly-convex functions (e.g. Hazan et al. [2007], Shalev-Shwartz and Singer [2007]). A snag is that the main analysis on the instantaneous regret term from these papers is identical to that of Lemma 3.2. Therefore, it appears as if more work is needed to derive an accelerated rate for linear coupling in the strongly-convex setting. We did not spend much time exploring this, but it seems that it should work out given enough time.

References

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *CoRR*, abs/1407.1537, 2014. URL http://arxiv.org/abs/1407.1537.

Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. 2015. URL http://arxiv.org/abs/1506.08187.

Michael Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. In *ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1018–1027. PMLR, 2018.

Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2019.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2845–2853. 2015.

Hongzhou Lin, Julien Mairal, and Zaïd Harchaoui. A universal catalyst for first-order optimization. In *NIPS 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3384–3392, 2015.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. In Jennifer G. Dy and Andreas Krause, editors, *ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80, pages 3331–3340, 2018.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269:543–547, 1983.

Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. ISBN 1-4020-7553-7.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Shai Shalev-Shwartz and Yoram Singer. Logarithmic regret algorithms for strongly convex repeated games. *The Hebrew University*, 2007.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2510–2518. 2014.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

A Proof of Lemma 3.3

Proof of Lemma 3.3. Let $k \geq 0$ and define $\gamma := \eta(1 - \eta L\rho/2)$. From Lemma 3.1 we have

$$\|\nabla f(x_{k+1})\|^2 \leq \frac{1}{\gamma}\mathbb{E}_k[f(x_{k+1}) - f(y_{k+1})] + \frac{L\eta^2}{2\gamma}\sigma.$$

Define $\Theta_k := D_w(x^*, z_k) - \mathbb{E}_k[D_w(x^*, z_{k+1})]$. Starting from the mirror descent guarantee given by Lemma 3.2 and setting $C := \rho/2\gamma = \frac{\rho}{2\eta(1-\eta L\rho/2)}$, we have

$$\begin{aligned}
&\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - x^*\rangle \\
&\leq \frac{\alpha_{k+1}^2}{2}\mathbb{E}_k[\|\nabla f(x_{k+1}, u_{k+1})\|^2] + \Theta_k \\
&\leq \frac{\alpha_{k+1}^2\rho}{2}\|\nabla f(x_{k+1})\|^2 + \frac{\alpha_{k+1}^2\sigma}{2} + \Theta_k \\
&\leq \frac{\alpha_{k+1}^2\rho}{2\gamma}\Big(\mathbb{E}_k[f(x_{k+1}) - f(y_{k+1})] + \frac{L\eta^2}{2}\sigma\Big) + \frac{\alpha_{k+1}^2\sigma}{2} + \Theta_k \\
&= \alpha_{k+1}^2 C\mathbb{E}_k[f(x_{k+1}) - f(y_{k+1})] + \frac{\alpha_{k+1}^2\sigma}{2}\Big(CL\eta^2 + 1\Big) + \Theta_k.
\end{aligned}$$

Re-arranging and using that $\mathbb{E}_k[f(x_{k+1})] = f(x_{k+1})$ (since x_{k+1} depends only of the randomness up to iteration k), we get

$$\begin{aligned}
\alpha_{k+1}^2 C\mathbb{E}_k[f(y_{k+1})] \leq \alpha_{k+1}^2 C f(x_{k+1}) + \frac{\alpha_{k+1}^2\sigma}{2}\Big(CL\eta^2 + 1\Big) \\
+ \Theta_k + \alpha_{k+1}\langle \nabla f(x_{k+1}), x^* - z_k\rangle.
\end{aligned} \tag{A.1}$$

This is the key point of the coupling of the mirror and gradient descent steps, and it shows us the value we need to put into τ_k so that we get the desired bound. From the definition of the iterates in Algorithm 1, we have $x_{k+1} = \tau_k z_k + (1-\tau_k)y_k$. Re-arranging gives $z_k = x_{k+1} + \frac{1-\tau_k}{\tau_k}(x_{k+1} - y_k)$. This together with the gradient inequality from convexity yields

$$\begin{aligned}
\alpha_{k+1}\langle \nabla f(x_{k+1}), x^* - z_k\rangle &= \alpha_{k+1}\frac{1-\tau_k}{\tau_k}\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}), x^* - x_{k+1}\rangle \\
&\leq \alpha_{k+1}\frac{1-\tau_k}{\tau_k}(f(y_k) - f(x_{k+1})) + \alpha_{k+1}(f(x^*) - f(x_{k+1})) \\
&= \Big(\frac{\alpha_{k+1}}{\tau_k} - \alpha_{k+1}\Big)(f(y_k) - f(x_{k+1})) + \alpha_{k+1}(f(x^*) - f(x_{k+1})) \\
&= (\alpha_{k+1}^2 C - \alpha_{k+1})(f(y_k) - f(x_{k+1})) + \alpha_{k+1}(f(x^*) - f(x_{k+1})) \\
&= (\alpha_{k+1}^2 C - \alpha_{k+1})f(y_k) - \alpha_{k+1}^2 C f(x_{k+1}) + \alpha_{k+1}f(x^*),
\end{aligned}$$

where in the second to last step we used that τ_k is defined in a way such that $\alpha_{k+1}/\tau_k = \alpha_{k+1}^2 C$. Finally, plugging the above into (A.1) yields the desired inequality. □