Convergence Analysis of Zeroth-Order Online Alternating Direction Method of Multipliers

Sijia Liu lsjxjtu@umich.edu
University of Michigan, Ann Arbor, MI 48109, USA
Pin-Yu Chen pin-yu.chen@ibm.com
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
Jie Chen dr.jie.chen@ieee.org
Northwestern Polytechnical University, Xi'an 710072, China
Alfred O. Hero hero@umich.edu
University of Michigan, Ann Arbor, MI 48109, USA

Abstract

In this paper, we design and analyze a new zeroth-order online algorithm, the zeroth-order online alternating direction method of multipliers (ZOO-ADMM). We prove that ZOO-ADMM has a convergence rate of $O(\sqrt{m}/\sqrt{T})$, where m is the number of optimization variables, and T is the number of iterations. Compared to the first-order gradient-based online algorithm, ZOO-ADMM requires $\sqrt{m}$ times more iterations, however, it enjoys dual advantages of being gradient-free operation and employing the ADMM to accommodate complex structured regularizers.

1 Introduction

Online convex optimization (OCO) performs sequential inference in a data-driven adaptive fashion, and has found a wide range of applications [1, 2, 3]. Several OCO algorithms have been proposed for regularized optimization, e.g., composite mirror descent, namely, proximal stochastic gradient descent [4], regularized dual averaging [5], and adaptive gradient descent [6]. However, the complexity of the aforementioned algorithms is dominated by the computation of the proximal operation with respect to the regularizers [7]. An alternative is to use online alternating direction method of multipliers (O-ADMM) [8, 9, 10]. Different from the algorithms in [4, 5, 6], the ADMM framework offers the possibility of splitting the optimization problem into a sequence of easily-solved subproblems. It was shown in [8, 9, 10] that the online variant of ADMM has convergence rate of $O(1/\sqrt{T})$ for convex loss functions and $O(\log T/T)$ for strongly convex loss functions, where T is the number of iterations.

One limitation of existing O-ADMM algorithms is the need to compute and repeatedly evaluate the gradient of the loss function over the iterations. In many practical scenarios, an explicit expression for the gradient is difficult to obtain. Examples are bandit optimization [11], simulation-based optimization problems [12, 13], and adversarial black-box machine learning models [14]. Moreover, in some high dimensional settings, acquiring the gradient information may be difficult, e.g., involving matrix inversion [15]. This motivates the development of gradient-free (zeroth-order) optimization algorithms.

Zeroth-order optimization approximates the full gradient via a randomized gradient estimate [11, 16, 17, 18, 19, 20]. For example, in [11, 19], zeroth-order algorithms were developed for bandit convex optimization with multi-point bandit feedback. In [16], a zeroth-order gradient descent algorithm was proposed that has $O(m/\sqrt{T})$ convergence rate, where m is the number of variables in the objective function. This slowdown in convergence rate was improved to $O(\sqrt{m}/\sqrt{T})$ in [17]. Its optimality was further proved in [18] under the framework of mirror descent algorithms.

Contributions: Different from the aforementioned zeroth-order algorithms, we design a new zeroth-order online ADMM (called ZOO-ADMM) algorithm, which enjoys advantages of gradient-free computation as well as ADMM. We prove that ZOO-ADMM yields a $O(\sqrt{m}/\sqrt{T})$ convergence rate for smooth+nonsmooth composite objective functions, which is at least as fast as existing zeroth-order algorithms in [17, 18, 19].

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

2 Preliminaries and Problem Formulation

We consider the following regularized loss minimization problem over a time horizon of length T:

$$\underset{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}}{\text{minimize}} \quad \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x};\mathbf{w}_t) + \phi(\mathbf{y}) \quad \text{subject to} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} = \mathbf{c}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^d$ are optimization variables, $\mathcal{X}$ and $\mathcal{Y}$ are closed convex sets, $f(\cdot;\mathbf{w}_t)$ is a convex and smooth cost/loss function parameterized by $\mathbf{w}_t$ at time t, ϕ is a convex regularization function (possibly nonsmooth), and $\mathbf{A} \in \mathbb{R}^{l\times m}$, $\mathbf{B} \in \mathbb{R}^{l\times d}$, and $\mathbf{c} \in \mathbb{R}^l$ are constraint coefficients associated with a system of l linear constraints. In problem (1), the use of time-varying cost functions $\{f(\mathbf{x};\mathbf{w}_t)\}_{t=1}^T$ captures environmental uncertainties that may exist in the online setting [1, 21]. One interpretation of $\{f(\mathbf{x};\mathbf{w}_t)\}_{t=1}^T$ is the empirical approximation to the stochastic objective function $\mathbb{E}_{\mathbf{w}\sim P}\left[f(\mathbf{x};\mathbf{w})\right]$. Here P is an empirical distribution with density $\sum_t \delta(\mathbf{w},\mathbf{w}_t)$, where $\{\mathbf{w}_t\}_{t=1}^T$ is a set of i.i.d. samples, and $\delta(\cdot,\mathbf{w}_t)$ is the Dirac delta function at $\mathbf{w}_t$. We also note that when $\mathcal{Y} = \mathcal{X}$, $l = m$, $\mathbf{A} = \mathbf{I}_m$, $\mathbf{B} = -\mathbf{I}_m$, $\mathbf{c} = \mathbf{0}_m$, the variable $\mathbf{y}$ and the linear constraint in (1) can be eliminated, leading to a standard OCO formulation. Here $\mathbf{I}_m$ denotes the $m \times m$ identity matrix, and $\mathbf{0}_m$ is the $m \times 1$ vector of all zeros.

To solve (1), a widely-used algorithm was developed by [8], which combines online proximal gradient descent and ADMM in the following form:

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\mathcal{X}}{\arg\min}\left\{\mathbf{g}_t^T\mathbf{x} - \boldsymbol{\lambda}_t^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}_t - \mathbf{c}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}_t - \mathbf{c}\|_2^2 + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|_{\mathbf{G}_t}^2\right\}, \tag{2}$$

$$\mathbf{y}_{t+1} = \underset{\mathbf{y}\in\mathcal{Y}}{\arg\min}\left\{\phi(\mathbf{y}) - \boldsymbol{\lambda}_t^T(\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{y} - \mathbf{c}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{y} - \mathbf{c}\|_2^2\right\}, \tag{3}$$

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t - \rho(\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{y}_{t+1} - \mathbf{c}), \tag{4}$$

where t is the iteration number (possibly the same as the time step), $\mathbf{g}_t$ is the gradient of the cost function $f(\mathbf{x};\mathbf{w}_t)$ at $\mathbf{x}_t$, namely, $\mathbf{g}_t = \nabla_{\mathbf{x}} f(\mathbf{x};\mathbf{w}_t)|_{\mathbf{x}=\mathbf{x}_t}$, $\boldsymbol{\lambda}_t$ is the Lagrangian multiplier (also known as the dual variable), ρ is a positive weight to penalize the augmented term associated with the equality constraint of (1), $\|\cdot\|_2$ denotes the ℓ_2 norm, η_t is a non-increasing sequence of positive step sizes, and $\|\mathbf{x} - \mathbf{x}_t\|_{\mathbf{G}_t}^2 = (\mathbf{x} - \mathbf{x}_t)^T\mathbf{G}_t(\mathbf{x} - \mathbf{x}_t)$ is a Bregman divergence generated by the strongly convex function $(1/2)\mathbf{x}^T\mathbf{G}_t\mathbf{x}$ with a known symmetric positive definite coefficient matrix $\mathbf{G}_t$.

To avoid explicit gradient calculations in (2), we adopt a randomized gradient estimator to estimate the gradient of a smooth cost function [16, 17, 18, 19]. The gradient estimate of $f(\mathbf{w};\mathbf{w}_t)$ is given by

$$\hat{\mathbf{g}}_t = \frac{f(\mathbf{x}_t + \beta_t\mathbf{z}_t;\mathbf{w}_t) - f(\mathbf{x}_t;\mathbf{w}_t)}{\beta_t}\mathbf{z}_t, \tag{5}$$

where $\mathbf{z}_t \in \mathbb{R}^m$ is a random vector drawn independently at each t from a distribution $\mathbf{z} \sim \mu$ with $\mathbb{E}_\mu[\mathbf{z}\mathbf{z}^T] = \mathbf{I}$, and $\{\beta_t\}$ is a non-increasing sequence of small positive smoothing constants. The rationale behind the estimator (5) is that $\hat{\mathbf{g}}_t$ becomes an unbiased estimator of $\mathbf{g}_t$ when the smoothing parameter β_t approaches zero [18].

This zeroth-order extension of O-ADMM (ZOO-ADMM) involves a modification of step (2) by replacing $\mathbf{g}_t$ with $\hat{\mathbf{g}}_t$:

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\mathcal{X}}{\arg\min}\left\{\hat{\mathbf{g}}_t^T\mathbf{x} - \boldsymbol{\lambda}_t^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}_t - \mathbf{c}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}_t - \mathbf{c}\|_2^2 + \frac{1}{2\eta_t}\|\mathbf{x} - \mathbf{x}_t\|_{\mathbf{G}_t}^2\right\}. \tag{6}$$

In (6), we can specify the matrix $\mathbf{G}_t$ in such a way as to cancel the term $\|\mathbf{A}\mathbf{x}\|_2^2$. This technique has been used in the linearized ADMM algorithms [7, 22] to avoid matrix inversions. Defining $\mathbf{G}_t = \alpha\mathbf{I} - \rho\eta_t\mathbf{A}^T\mathbf{A}$, the update rule (6) simplifies to a projection operator

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}\in\mathcal{X}}{\arg\min}\left\{\|\mathbf{x} - \boldsymbol{\omega}\|_2^2\right\};\ \ \boldsymbol{\omega} := \left[\frac{\eta_t}{\alpha}\left(-\mathbf{g}_t + \mathbf{A}^T(\boldsymbol{\lambda}_t - \rho\mathbf{A}\mathbf{x}_t - \rho\mathbf{B}\mathbf{y}_t + \rho\mathbf{c})\right) + \mathbf{x}_t\right], \tag{7}$$

where $\alpha > 0$ is a parameter such that $\mathbf{G}_t \succeq \mathbf{I}$, and $\mathbf{X} \succeq \mathbf{Y}$ means $\mathbf{X} - \mathbf{Y}$ is positive semidefinite.

To evaluate the convergence behavior of ZOO-ADMM, we will derive its expected average regret [1]

$$\overline{\text{Regret}}_T(\mathbf{x}_t,\mathbf{y}_t) := \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}_t;\mathbf{w}_t) + \phi(\mathbf{y}_t)\right) - \frac{1}{T}\sum_{t=1}^{T}\left(f(\mathbf{x}^*;\mathbf{w}_t) + \phi(\mathbf{y}^*)\right)\right], \tag{8}$$

where $(\mathbf{x}^*,\mathbf{y}^*)$ denotes the best batch offline solution. Note that the alternating structure in ZOO-ADMM requires a different regret analysis compared to existing zeroth-order algorithms [16, 17, 18].

Algorithm 1 ZOO-ADMM for solving problem (1)

```
Input: $\mathbf{x}_1 \in \mathcal{X}$, $\mathbf{y}_1 \in \mathcal{Y}$, $\boldsymbol{\lambda}_1 = \mathbf{0}$, $\rho > 0$, step sizes $\{\eta_t\}$, smoothing constants $\{\beta_t\}$, distribution $\mu$, and
    $\alpha \geq \rho\eta_t\lambda_{\max}(\mathbf{A}^T\mathbf{A}) + 1$ so that $\mathbf{G}_t \succeq \mathbf{I}$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a symmetric
    matrix
for $t = 1, 2, \ldots, T$ do
    sample $\mathbf{z}_t \sim \mu$ to generate $\hat{\mathbf{g}}_t$ using (5)
    update $\mathbf{x}_{t+1}$ via (7) under $\hat{\mathbf{g}}_t$ and $(\mathbf{x}_t, \mathbf{y}_t, \boldsymbol{\lambda}_t)$
    update $\mathbf{y}_{t+1}$ via (3) under $(\mathbf{x}_{t+1}, \boldsymbol{\lambda}_t)$
    update $\boldsymbol{\lambda}_{t+1}$ via (4) under $(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \boldsymbol{\lambda}_t)$
    if $\mathbf{B}$ is invertible then
        compute $\mathbf{y}'_{t+1} := \mathbf{B}^{-1}(\mathbf{c} - \mathbf{A}\mathbf{x}_{t+1})$
    else
        compute $\mathbf{x}'_{t+1} := \mathbf{A}^{-1}(\mathbf{c} - \mathbf{B}\mathbf{y}_{t+1})$
    end if
end for
output: $\{\mathbf{x}_t, \mathbf{y}'_t\}$ or $\{\mathbf{x}'_t, \mathbf{y}_t\}$.
```

3 ZOO-ADMM and Convergence Analysis

Algorithm 1 summarizes the procedures of the proposed ZOO-ADMM for solving problem (1). We next present the convergence analysis of ZOO-ADMM under the following assumptions:
• *Assumption A:* In problem (1), $\mathcal{X}$ and $\mathcal{Y}$ are bounded with finite diameter R, and at least one of $\mathbf{A}$ and $\mathbf{B}$ is invertible.

• *Assumption B:* $f(\cdot;\mathbf{w}_t)$ is convex and Lipschitz continuous with $\sqrt{\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x};\mathbf{w}_t)\|_2^2]} \leq L_1$ for all t and $\mathbf{x} \in \mathcal{X}$.

• Assumption C: $f(\cdot;\mathbf{w}_t)$ is $L_g(\mathbf{w}_t)$-smooth with $L_g = \sqrt{\mathbb{E}[(L_g(\mathbf{w}_t)^2)]}$.

• *Assumption D:* ϕ is convex and L_2-Lipschitz continuous with $\|\partial\phi(\mathbf{y})\|_2 \leq L_2$ for all $\mathbf{y} \in \mathcal{Y}$, where $\partial\phi(\mathbf{y})$ denotes the subgradient of ϕ.

• *Assumption E:* In (5), given $\mathbf{z} \sim \mu$, the quantity $M(\mu) := \sqrt{\mathbb{E}[\|\mathbf{z}\|_2^6]}$ is finite, and there is a function $s : \mathbb{N} \to \mathbb{R}_+$ satisfying $\mathbb{E}[\|\langle \mathbf{a}, \mathbf{z}\rangle \mathbf{z}\|_2^2] \leq s(m)\|\mathbf{a}\|_2^2$ for all $\mathbf{a} \in \mathbb{R}^m$, where $\langle\cdot,\cdot\rangle$ denotes the inner product of two vectors.

We remark that Assumptions A-D are standard for stochastic gradient-based and ADMM-type methods [1, 8, 21, 23]. Assumption E places moment constraints on the distribution μ that will allow us to derive the necessary concentration bounds for our convergence analysis. If μ is uniform on the surface of the Euclidean-ball of radius $\sqrt{m}$, we have $M(\mu) = m^{1.5}$ and $s(m) = m$. And if $\mu = \mathcal{N}(\mathbf{0}, \mathbf{I}_{m\times m})$, we have $M(\mu) \approx m^{1.5}$ and $s(m) \approx m$ [18]. For ease of representation, we restrict attention to the case that $s(m) = m$ in the rest of the paper.

Theorem 1 *Suppose* $\mathbf{B}$ *is invertible in problem (1). For* $\{\mathbf{x}_t, \mathbf{y}'_t\}$ *generated by ZOO-ADMM, the expected average regret is bounded as*

$$\overline{\mathrm{Regret}}_T(\mathbf{x}_t, \mathbf{y}'_t) \leq \frac{1}{T}\sum_{t=2}^{T}\left\{\frac{\alpha}{2\eta_t} - \frac{\alpha}{2\eta_{t-1}} - \frac{\sigma}{2}, 0\right\} R^2 + \frac{mL_1^2}{T}\sum_{t=1}^{T}\eta_t + \frac{M(\mu)^2 L_g^2}{4T}\sum_{t=1}^{T}\eta_t\beta_t^2 + \frac{K}{T}, \quad (9)$$

where α *is introduced in (7),* R*,* L_1*,* L_g*,* $s(m)$ *and* $M(\mu)$ *are defined in Assumptions A-E, and* K *denotes a constant term that depends on* α*,* R*,* η_1*,* $\mathbf{A}$*,* $\mathbf{B}$*,* $\boldsymbol{\lambda}$*,* ρ *and* L_2*. Suppose* $\mathbf{A}$ *is invertible in (1). For* $\{\mathbf{x}'_t, \mathbf{y}_t\}$*, the regret* $\overline{\mathrm{Regret}}_T(\mathbf{x}'_t, \mathbf{y}_t)$ *obeys the same bounds as (9).*

Proof: [24, Appendix A]. ■

In Theorem 1, if the step size η_t and the smoothing parameter β_t are chosen by $\eta_t = \frac{C_1}{m\sqrt{t}}$ and $\beta_t = \frac{C_2}{M(\mu)t}$ for some constant $C_1 > 0$ and $C_2 > 0$, then the regret bound (9) can simplify to

$$\overline{\mathrm{Regret}}_T(\mathbf{x}_t, \mathbf{y}'_t, \mathbf{x}^*, \mathbf{y}^*) \leq \frac{\alpha R^2}{2C_1}\frac{\sqrt{m}}{\sqrt{T}} + 2C_1 L_1^2\frac{\sqrt{m}}{\sqrt{T}} + \frac{5C_1C_2^2L_g^2}{12}\frac{1}{T} + \frac{K}{T}. \quad (10)$$

It is clear from (10) that ZOO-ADMM converges at least as fast as $O(\sqrt{m}/\sqrt{T})$, which is similar to the convergence rate of O-ADMM found by [8] but involves an additional factor $\sqrt{m}$. Such a dimension-dependent effect on the convergence rate has also been reported for other zeroth-order optimization algorithms [17, 18, 19], leading to the same convergence rate as ours.

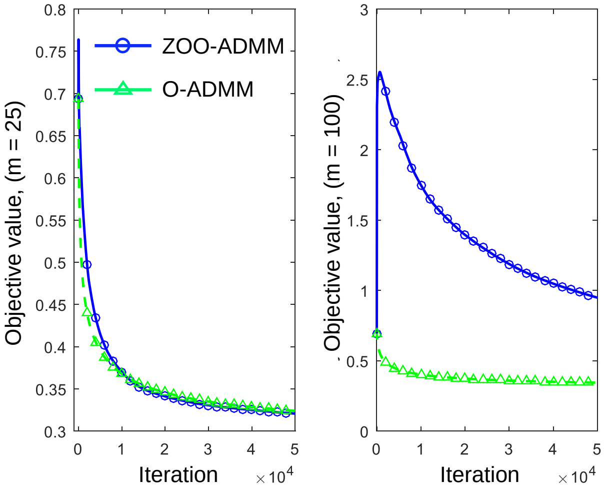

Figure 1: Objective value of overlapping group Lasso versus ZOO-ADMM iteration for $m = 25$ (left) and $m = 100$ (right).

For demonstration, we compare ZOO-ADMM with the conventional O-ADMM algorithm in [8] with the same parameters: $\mathbf{x}_1 = \mathbf{0}$, $\mathbf{y}_1 = \mathbf{0}$, $\boldsymbol{\lambda}_1 = \mathbf{0}$, $\rho = 10$, $\eta_t = 1/\sqrt{mt}$, $\beta_t = 1/(m^{1.5}t)$, and $\alpha = \rho\eta_t\lambda_{\max}(\mathbf{A}^T\mathbf{A})+1$. The distribution μ is chosen to be uniform on the surface of the Euclidean-ball of radius $\sqrt{m}$. We consider a simulated sparse classification task [8], where an overlapping group Lasso regularization is imposed.

Similar to [8], we draw $n = 512$ random samples $\{\mathbf{a}_i\}$ from $\mathcal{N}(\mathbf{0}_m, \mathbf{I}_m)$, and we generate $\{c_i\}$ as $c_i = \operatorname{sign}(\mathbf{a}_i^T\mathbf{x}^* + \epsilon_i)$, where $\mathbf{x}^*$ is a given coefficient vector, $\epsilon_i \in \mathcal{N}(\mathbf{0}, 0.01\mathbf{I})$ denotes sample noise, and $\operatorname{sign}(x) = 1$ if $x \geq 0$ and 0 otherwise. Here $\{(\mathbf{a}_i, c_i) : \mathbf{a}_i \in \mathbb{R}^m, c_i \in \{-1, 1\}\}_{i=1}^n$ gives a set of training samples in which $\mathbf{a}_i$ is the input feature vector with dimension $m = \tilde{m}^2$ for some $\tilde{m}$, and c_i is the target (output) variable. In (1), we design a logistic regression classifier with loss function $f_t(\mathbf{x}) = \log(1 + e^{-c_t\mathbf{a}_t^T\mathbf{x}})$, where $\mathbf{x}$ is the coefficients to be designed, and $\mathbf{w}_t = (\mathbf{a}_t, c_t)$. To promote group sparsity, we convert a coefficient vector $\mathbf{x} \in \mathbb{R}^m$ into a $\tilde{m} \times \tilde{m}$ matrix, denoted by $\mathbf{X}$, and impose column-wise and row-wise group sparsity via the regularizer $\phi(\mathbf{x}) = \sum_{i=1}^{\tilde{m}} (\|\mathbf{X}_{i,\cdot}\|_2 + \|\mathbf{X}_{\cdot,i}\|_2)$.

Fig. 1 presents the convergence trajectory of ZOO-ADMM as a function of number of iterations for problem dimension $m \in \{25, 100\}$. As we can see, the convergence speed of ZOO-ADMM is comparable to that of O-ADMM when m is small. However, the convergence becomes much slower when a larger-scale problem is considered. The dependency of the convergence rate of ZOO-ADMM on the dimension of optimization variables is precisely characterized by Theorem 1.

4 Conclusion an Future Work

We proposed and analyzed a new zeroth-order online ADMM algorithm, ZOO-ADMM. We showed that the regret bound of ZOO-ADMM suffers an additional dimension-dependent factor in convergence rate over gradient-based online variants of ADMM, leading to $O(\sqrt{m}/\sqrt{T})$ convergence rate, where m is the number of optimization variables. In the future, we would like to analyze the converge of ZOO-ADMM with variance reduction techniques, e.g., minibatch sampling. We also would like to relax the assumptions on smoothness and convexity of the cost function in ZOO-ADMM.

Acknowledgements

The work of S. Liu and A. O. Hero was partially supported by grants from the US Army Research Office, grant numbers W911NF-15-1-0479 and W911NF-15-1-0241. The work of J. Chen was supported in part by the Natural Science Foundation of China under Grant 61671382.

References

[1] E. Hazan, “Introduction to online convex optimization,” *Foundations and Trends® in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[2] S. Hosseini, A. Chapman, and M. Mesbahi, “Online distributed convex optimization on dynamic networks,” *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3545–3550, 2016.

[3] E. C. Hall and R. M. Willett, “Online convex optimization in dynamic environments,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 4, pp. 647–662, June 2015.

[4] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, “Composite objective mirror descent.” in *COLT*, 2010, pp. 14–26.

[5] L. Xiao, “Dual averaging methods for regularized stochastic learning and online optimization,” *Journal of Machine Learning Research*, vol. 11, no. Oct., pp. 2543–2596, 2010.

[6] J. Duchi, E. Hazan, and Y. Singer, “Adaptive subgradient methods for online learning and stochastic optimization,” *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[7] N. Parikh and S. Boyd, “Proximal algorithms,” *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[8] T. Suzuki, “Dual averaging and proximal gradient descent for online alternating direction multiplier method,” in *International Conference on Machine Learning*, 2013, pp. 392–400.

[9] H. Ouyang, N. He, L. Tran, and A. Gray, “Stochastic alternating direction method of multipliers,” in *International Conference on Machine Learning*, 2013, pp. 80–88.

[10] H. Wang and A. Banerjee, “Online alternating direction method (longer version),” *arXiv preprint arXiv:1306.3721*, 2013.

[11] A. Agarwal, O. Dekel, and L. Xiao, “Optimal algorithms for online convex optimization with multi-point bandit feedback.” in *COLT*, 2010, pp. 28–40.

[12] J. C. Spall, *Introduction to stochastic search and optimization: estimation, simulation, and control.* John Wiley & Sons, 2005, vol. 65.

[13] L. M. Rios and N. V. Sahinidis, “Derivative-free optimization: a review of algorithms and comparison of software implementations,” *Journal of Global Optimization*, vol. 56, no. 3, pp. 1247–1293, 2013.

[14] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, “Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models,” *arXiv preprint arXiv:1708.03999*, 2017.

[15] S. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge university press, 2004.

[16] Y. Nesterov and V. Spokoiny, “Random gradient-free minimization of convex functions,” *Foundations of Computational Mathematics*, vol. 2, no. 17, pp. 527–566, 2015.

[17] S. Ghadimi and G. Lan, “Stochastic first-and zeroth-order methods for nonconvex stochastic programming,” *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368, 2013.

[18] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, “Optimal rates for zero-order convex optimization: The power of two function evaluations,” *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[19] O. Shamir, “An optimal algorithm for bandit and zero-order convex optimization with two-point feedback,” *Journal of Machine Learning Research*, vol. 18, no. 52, pp. 1–11, 2017.

[20] D. Hajinezhad, M. Hong, and A. Garcia, “Zenith: A zeroth-order distributed algorithm for multi-agent nonconvex optimization,” 2017.

[21] S. Shalev-Shwartz, “Online learning and online convex optimization,” *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2012.

[22] X. Zhang, M. Burger, and S. Osher, “A unified primal-dual algorithm framework based on bregman iteration,” *Journal of Scientific Computing*, vol. 46, no. 1, pp. 20–46, 2011.

[23] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[24] S. Liu, J. Chen, P.-Y. Chen, and A. O. Hero, “Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications,” *arXiv preprint arXiv:1710.07804*, 2017.