Explicit Regularization of Stochastic Gradient Methods through Duality

Anant Raj
MPI for Intelligent Systems, Tübingen

Francis Bach
Inria, Ecole Normale Supérieure
PSL Research University, Paris, France.

Abstract

We consider stochastic gradient methods under the interpolation regime where a perfect fit can be obtained (minimum loss at each observation). While previous work highlighted the *implicit* regularization of such algorithms, we consider an *explicit* regularization framework as a minimum Bregman divergence convex feasibility problem. Using convex duality, we propose randomized Dykstra-style algorithms based on randomized dual coordinate ascent. For non-accelerated coordinate descent, we obtain an algorithm which bears strong similarities with (non-averaged) stochastic mirror descent on specific functions, as it is equivalent for quadratic objectives, and equivalent in the early iterations for more general objectives. It comes with the benefit of an explicit convergence theorem to a minimum norm solution. For accelerated coordinate descent, we obtain a new algorithm that has better convergence properties than existing stochastic gradient methods in the interpolating regime. This leads to accelerated versions of the perceptron for generic ℓ_p-norm regularizers, which we illustrate in experiments.

1 Introduction

With the recent advancement in machine learning and hardware research, the size and capacity of training models for machine learning tasks have been consistently increasing. For many models which are widely used in practice, e.g., deep neural networks (Goodfellow et al., 2016) and non-parametric regression models (Belkin et al., 2018; Liang & Rakhlin, 2018), the training process achieves zero error, which means that such models are expressive enough to interpolate the training data completely. Hence, it is important to understand the interpolation regime to improve the training and prediction of such complex and over-parameterized models used in machine learning.

It is a well known fact that regularization, either explicit or implicit, plays a crucial role in achieving better generalization. While Tikhonov regularization is amongst the most famous form of regularization (Golub et al., 1999; Weese, 1993) for linear or non-linear problems, several other methods can induce regularization in form of computational regularization when training machine learning models (Yao et al., 2007; Rudi et al., 2015; Srivastava et al., 2014). Apart from explicitly induced regularization in machine learning models, optimization algorithms like (stochastic) gradient descent which is widely used in practice while training large machine learning models, also induce implicit regularization in the obtained solution. In many cases, (stochastic) gradient descent converges to minimum Euclidean norm solutions. Recent series of papers (Soudry et al., 2018; Gunasekar et al., 2018; Kubo et al., 2019; Arora et al., 2019) present result about introducing implicit regularization/bias by (stochastic) gradient descent in different set of convex and non-convex problems.

In this paper, we address the following question: instead of relying on implicit regularization properties of stochastic algorithms, can we introduce an *explicit* regularization/bias while training over-parameterized models in the interpolation regime?

In optimization terms, the interpolation regime corresponds to the minimization of an average of finitely many functions of the form

$$F(\theta) = \frac{1}{n}\sum_{i=1}^{n} f_i(\theta),$$

with respect to $\theta \in \mathbb{R}^d$, where there is a global minimizer of F, which happens to be a global minimizer of

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

all functions f_i, for $i \in \{1, \dots, n\}$ (instead of only minimizing their average). In the interpolation regime, we are thus looking for a point $\theta \in \mathbb{R}^d$ in the intersection of all sets of minimizers

$$\mathcal{K}_i = \arg\min_{\eta \in \mathbb{R}^d} f_i(\eta),$$

for all $i \in \{1, \dots, n\}$.

We can thus explicitly regularize the problem by solving the following optimization problem:

$$\min_{\theta \in \mathbb{R}^d} \psi(\theta) \text{ such that } \forall i \in \{1, \dots, n\},\ \theta \in \mathcal{K}_i, \quad (1)$$

where ψ is a regularization function (typically a squared norm). In the reformulated problem given in Eq. (1), explicit regularization can be induced in the solution via the structure of the function ψ. Note also that the above problem can be seen as problem of generalized projection onto sets, which are convex if the original functions f_i's are convex, which we assume throughout this paper.

To address the problem defined in Eq. (1), we use the tools from convex duality and accelerated randomized coordinate ascent, which result in Dykstra-style projection algorithms (Boyle & Dykstra, 1986; Zhang et al., 2008; Gaffke & Mathar, 1989). In this paper, we make the following contributions:

(a) We provide a generic inequality going from dual guarantees in function values to primal guarantees in terms of Bregman divergences of iterates.

(b) For non-accelerated coordinate ascent, we obtain an algorithm which bears strong similarities with (non-averaged) stochastic mirror descent on specific functions f_i's. Our algorithm comes with the benefit of an explicit convergence theorem to a minimum value of the regularizer.

(c) For accelerated coordinate ascent, we obtain a new algorithm that has better convergence properties than existing stochastic gradient methods in the interpolating regime. While we indeed use the classical accelerated randomized coordinate descent algorithm to get accelerated rates, we show that we do not need any of the strong assumption that previous attempts at acceleration were needing (e.g.,Vaswani et al. (2018)) for SGD in interpolation regime.

(d) This leads to accelerated versions of the perceptron for generic ℓ_p-norm regularizers (this is already an improvement for the ℓ_2-regularizer).

1.1 Related work

Stochastic gradient methods. First order stochastic gradient based iterative approaches (Nemirovski et al., 2009; Duchi et al., 2011; Kingma & Ba, 2014; Defazio et al., 2014; Ward et al., 2019) are the most efficient methods to perform optimization for machine learning problems with large datasets. There has been a large amount of work done in the area of stochastic first order optimization methods (see, e.g., Polyak, 1990; Polyak & Juditsky, 1992; Nemirovski et al., 2009; Bach & Moulines, 2011, and references therein) since the original stochastic approximation approach was proposed by Robbins & Monro (1951).

Primal SGD in the interpolation regime. To address the optimization problem in the interpolation regime, Vaswani et al. (2018) provide faster convergence rates for first order stochastic methods in the Euclidean geometry. They propose a strong growth condition, and a more widely applicable weak growth condition, under which stochastic gradient descent algorithm achieves fast convergence rate while using constant learning rate (a side contribution of our paper is to extend the latter algorithm to stochastic mirror descent). Vaswani et al. (2019) propose to use line-search to set the step-size while training over-parameterized models which can fit completely to data. Several other works propose to use constant learning rate for stochastic gradient methods (Ma et al., 2017; Bassily et al., 2018; Liu & Belkin, 2018; Cevher & Vũ, 2019) while training extremely expressive models which interpolate. However, all of the above mentioned works are primal-based algorithms.

Dysktra's projection algorithms. Dykstra-type projection algorithms (Boyle & Dykstra, 1986; Gaffke & Mathar, 1989) are simple modifications of the classical alternating projections methods (Von Neumman, 1951; Halperin, 1962) to project on the intersection of convex sets. A key interpretation is the connection between Dykstra's algorithm and block coordinate ascent (Bauschke & Koch, 2015; Bauschke & Combettes, 2011; Tibshirani, 2017), which we use in this paper. Chambolle et al. (2017) provide accelerated rates for Dykstra projection algorithm when projecting on the intersection of two sets.

Coordinate descent. Coordinate descent has a long history in the optimization literature (Tseng & Bertsekas, 1987; Tseng, 1993, 2001). Rates for accelerated randomized coordinate descent were first proved by Nesterov (2012). Since then, various extensions of the accelerated coordinate descent including proximal accelerated coordinate descent and non-uniform sampling have been proposed by Lin et al. (2015); Allen-Zhu et al. (2016); Nesterov & Stich (2017); Hendrikx et al.

(2019). Dual coordinate ascent can also be used to solve regularized empirical risk minimization problem (Shalev-Shwartz & Zhang, 2013, 2014). We recover some of their results as a by-product in this paper.

Perceptron. The perceptron is one of the oldest machine learning algorithms (Block, 1962; Minsky & Papert, 2017). Since then, there has been a lot of work on theoretical and empirical foundations of perceptron algorithms (Freund & Schapire, 1999; Shalev-Shwartz & Singer, 2005; Tsampouka & Shawe-Taylor, 2005), in particular, with related extensions to ours, to ℓ_p-norm perceptron through mirror maps (Grove et al., 2001; Kivinen, 2003). However, none of the above mentioned work forces structure to the optimal solution in an explicit way.

2 Optimization Algorithms for Finite Data

We consider the finite data setting, that is, we will give bounds on training objectives (or distances to the minimum norm interpolator on the training set). We thus consider the problem:

$$\min_{\theta \in \mathbb{R}^d} \psi(\theta) \text{ such that } \forall i \in \{1, \dots, n\},\ x_i^\top \theta \in \mathcal{Y}_i, \quad (2)$$

where:

- Regularizer / mirror map: $\psi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a differentiable μ-strongly convex function with respect to some norm $\|\cdot\|$ (which is not in general the ℓ_2-norm). We will consider in this paper the associated Bregman divergence (Bregman, 1967) defined as

$$D_\psi(\theta, \eta) = \psi(\theta) - \psi(\eta) - \psi'(\eta)^\top(\theta - \eta).$$

- Data: $x_i \in \mathbb{R}^{d \times k}$, $\mathcal{Y}_i \subset \mathbb{R}^k$ are closed convex sets, for $i \in \{1, \dots, n\}$.
- Feasibility / interpolation regime: we make the assumption that there exists $\theta \in \mathbb{R}^d$ such that $\psi(\theta) < \infty$ and $\forall i \in \{1, \dots, n\},\ x_i^\top \theta \in \mathcal{Y}_i$.

This is a general formulation that includes any set $\mathcal{K}_i$ like in the introduction (by having $k = d$, $x_i = I$, and $\mathcal{Y}_i = \mathcal{K}_i$), with an important particular case $k = 1$ (classical linear prediction).

In this paper, we consider primarily the ℓ_p-norm set-up, where $\psi(\theta) = \frac{1}{2}\|\theta\|_p^2$ for $p \in (1, 2]$, which is $(p-1)$-strongly convex with respect to the ℓ_p-norm (Ball et al., 1994; Duchi et al., 2010). The simplex with the entropy mirror map, which is 1-strongly convex with respect to the ℓ_1-norm, could also be considered.

2.1 From dual guarantees to primal guarantees

We can use Fenchel duality to obtain a dual problem for the problem given in Eq.(2). We will need the support function $\sigma_{\mathcal{Y}_i}$ of the convex set $\mathcal{Y}_i$, defined as, for $\alpha_i \in \mathbb{R}^k$ (Boyd & Vandenberghe, 2004),

$$\sigma_{\mathcal{Y}_i}(\alpha_i) = \sup_{y_i \in \mathcal{Y}_i} y_i^\top \alpha_i.$$

We have, by Fenchel duality:

$$\min_{\theta \in \mathbb{R}^d} \psi(\theta) \text{ such that } \forall i \in \{1, \dots, n\},\ x_i^\top \theta \in \mathcal{Y}_i \quad (3)$$

$$= \min_{\theta \in \mathbb{R}^d} \psi(\theta) + \frac{1}{n} \sum_{i=1}^n \max_{\alpha_i \in \mathbb{R}^k} \Big\{ \alpha_i^\top x_i^\top \theta - \sigma_{\mathcal{Y}_i}(\alpha_i) \Big\}$$

$$= \max_{\forall i,\ \alpha_i \in \mathbb{R}^k} -\frac{1}{n} \sum_{i=1}^n \sigma_{\mathcal{Y}_i}(\alpha_i) - \psi^\star\Big(-\frac{1}{n} \sum_{i=1}^n x_i \alpha_i \Big), \quad (4)$$

with, at optimality,

$$\theta^\star = \theta(\alpha^\star) = \nabla \psi^\star\Big(-\frac{1}{n} \sum_{i=1}^n x_i \alpha_i \Big).$$

We denote by $G(\alpha)$ the dual objective function above. With our assumptions of feasibility and strong-convexity of ψ, there is a unique minimizer $\theta^\star \in \mathbb{R}^d$. The dual problem is bounded from above, and we assume that there exists a maximizer $\alpha^\star \in \mathbb{R}^{n \times k}$.

In this paper, we will consider dual algorithms to solve the problem disccused earlier in this section, that naturally leads to guarantees on the gap

$$\text{gap}(\alpha) = G(\alpha^\star) - G(\alpha).$$

Our first result is to provide some primal guarantees from $\theta(\alpha)$.

Proposition 1 *With our assumption, for any $\alpha \in \mathbb{R}^{n \times k}$, we have:*

$$D_\psi(\theta^\star, \theta(\alpha)) \leqslant \text{gap}(\alpha).$$

In the above statement, we also assume that ψ is differentiable everywhere, since Bregman divergences are well defined for differentiable functions. However, if we want to relax the above statement for a general function ψ which might not be differentiable, we would need to replace the term $D_\psi(\theta^\star, \theta(\alpha))$ in Eq. (1) with $\psi(\theta^\star) - \psi(\theta(\alpha))) - \langle \partial\psi(\theta(\alpha))), \theta^\star - \theta(\alpha)) \rangle$ where $\partial\psi(\theta(\alpha)))$ is a specific sub-gradient of ψ at point $\theta(\alpha)$. In the proof of Proposition 1, we simply use the duility structure of the problem with Fenchel-Young inequality. See the detailed proof in Appendix A.

This result relates primal rate of convergence and dual rate of convergence, and holds true irrespective of the algorithm used to optimize the dual objective. Using it, we can recover convergence guarantees for stochastic dual coordinate ascent (SDCA) (Shalev-Shwartz & Zhang, 2013) and accelerated SDCA (Shalev-Shwartz & Zhang, 2014). Compared to their analysis, our result directly provides rates of convergence from existing results in coordinate descent, but in terms of primal iterates. Details are provided in Appendix C.

Overall, we limit our discussion to convex functions however there is no requirement of the linear model to be used. Generalization of Proposition 1 (Proposition 2) holds for general convex objective and can be extended to non-linear models without extra effort.

2.2 Randomized coordinate descent

Given our relationship between primal iterate sub-optimality and dual sub-optimality gap gap(α) for any dual variable α and its corresponding primal variable $\theta(\alpha)$, we can leverage good existing algorithms on the dual problem. One such well known method is randomized dual coordinate descent, where α and thus $\theta(\alpha)$ will be random.

The algorithm is initialized with $\alpha_i^{(0)} = 0$ for all $i \in \{1, \dots, n\}$, and at step $t > 0$, an index $i(t) \in \{1, \dots, n\}$ is selected uniformly (for simplicity) at random. The update for proximal randomized coordinate ascent (Richtárik & Takáč, 2014) is obtained in the following lemma (whose proof is given in Appendix A.1).

Lemma 1 *For any uniformly randomly selected coordinate $i(t)$ at time instance t, the update for randomized proximal coordinate ascent is equal to*

$$\alpha_{i(t)} = \alpha_{i(t)}^{(t-1)} + \frac{n}{L_{i(t)}} x_{i(t)}^\top \theta(\alpha^{(t-1)}) - \frac{n}{L_{i(t)}} \Pi_{\mathcal{Y}_i}\Big(\frac{L_{i(t)}}{n}\alpha_{i(t)}^{(t-1)} + x_{i(t)}^\top \theta(\alpha^{(t-1)})\Big),$$

where $\Pi_{\mathcal{Y}_i}$ is the orthogonal projection on $\mathcal{Y}_i$, and L_i is equal to $L_i = \frac{1}{\mu}\|x_i\|^2_{2\to\star} = \frac{1}{\mu}\sup_{\|\beta_i\|_2=1}\|x_i\beta_i\|_\star^2$.

Here, we implicitly assume that the individual projections on convex set $\mathcal{Y}_i$ for all $i \in \{1, \cdots, n\}$ are easy to compute, leading to Algorithm 1. For uniformly random selection of the datapoint $x_{i(t)}$ at time t, $L_i(t)$ can simply be replaced by $\max_i L_i$ in the algorithm.

Algorithm 1 Proximal Random Coordinate Ascent

Input: α_0, $\theta_0 \leftarrow \theta(\alpha_0)$ and $\mathbf{x}_i, \mathcal{Y}_i$ for $i \in [n]$.
Output: θ_{T+1} and α_{T+1}
for $t \leftarrow 1$ **to** T **do**
  Choose $i(t) \in \{1, 2, \cdots, n\}$ randomly.
  $\beta_{(\text{prev})} = \alpha_{i(t)}^{(t-1)}$
  $\zeta_t = \Pi_{\mathcal{Y}_i}\Big(\frac{L_{i(t)}}{n}\alpha_{i(t)}^{(t-1)} + x_{i(t)}^\top\theta_{t-1}\Big)$.
  $\alpha_{i(t)}^{(t)} = \alpha_{i(t)}^{(t-1)} + \frac{n}{L_{i(t)}} x_{i(t)}^\top \theta_{t-1} - \frac{n}{L_{i(t)}}\zeta_t$.
  $\Delta_\beta = \alpha_{i(t)} - \beta_{(\text{prev})}$.
  Update $\theta_t \leftarrow \theta(\alpha_t)$ $\{$Use $\Delta_\beta, x_{i(t)}\}$.

Proximal randomized coordinate descent is a well studied problem (Nesterov & Stich, 2017; Richtárik & Takáč, 2014), and has a known rate of convergence for smooth objective functions. The set of optimal solutions of the dual problem in Equation (4) is denoted by $A^\star$ and $\alpha^\star$ is an element of it. Define,

$$\mathcal{R}(\alpha) = \max_y \max_{\alpha^\star \in A^\star} \{\|y - \alpha^\star\| \ : \ G(y) \geq G(\alpha)\}.$$

Since we assumed that ψ is μ-strongly convex, $\psi^\star$ is $(\frac{1}{\mu})$-smooth, and we get

$$\mathbb{E}\Big[D_\psi(\theta^\star, \theta(\alpha^{(t)}))\Big] \leqslant \mathbb{E}\big[\text{gap}(\alpha^{(t)})\big] \leqslant \frac{\max_i L_i}{t}\frac{\max\{\|\alpha^\star\|^2, \mathcal{R}(0)^2\}}{n}, \quad (5)$$

where L_i is defined in Lemma 1. The convergence rate given in Eq. (5) can further be improved with non-uniform sampling based on the values L_i, and then $\max_i L_i$ can be replaced by $\frac{1}{n}\sum_{i=1}^n L_i$ (Richtárik & Takáč, 2014).

2.3 Relationship to least-squares

We now discuss an important case of the above formulation when $\mathcal{Y}_i$ is a singleton set, i.e., $\mathcal{Y}_i = \{y_i\}$. This problem has been addressed recently by Calatroni et al. (2019) and we recover it as a special case of our general formulation.

We will make a link with least-squares in the interpolation regime, which can be written as a finite sum objective as follows,

$$\min \left[\frac{1}{2n}\sum_{i=1}^n \|y_i - x_i^\top\theta\|_2^2 = \frac{1}{2n}\sum_{i=1}^n d(x_i^\top\theta, \mathcal{Y}_i)^2\right]. \quad (6)$$

It turns out that primal stochastic mirror descent with constant step-size applied to Eq. (6) and our formulation provided in Section 2.1 are equivalent, as we now show.

Lemma 2 *Consider the stochastic mirror descent updates using the mirror map ψ for the least-squares problem provided in Eq. (6). Then, the corresponding stochastic mirror descent updates converges to minimum ψ solution.*

Proof Consider the primal-dual formulation given in Eq. (3) and Eq. (4), with $\mathcal{Y}_i = \{y_i\}$. The randomized dual coordinate ascent has the following update rule:

$$\alpha_{i(t)}^{(t)} = \alpha_{i(t)}^{(t-1)} + \frac{n}{L_{i(t)}}(x_{i(t)}^\top \theta(\alpha^{(t-1)}) - y_{i(t)}). \quad (7)$$

From the first order optimality condition, the update in Eq. (7) translates into, with $\theta^{(t)} = \theta(\alpha^{(t)})$,

$$\psi'(\theta^{(t)}) = \psi'(\theta^{(t-1)}) - \frac{1}{L_{i(t)}} x_{i(t)}(x_{i(t)}^\top \theta(\alpha^{(t-1)}) - y_{i(t)}),$$

which is exactly stochastic mirror descent on the least-squares objective with mirror map ψ. Hence the result. ■

The rate of convergence can be obtained by the use of Eq. (5).

General case (beyond singletons). For any set $\mathcal{Y}_i$, if $\alpha_{i(t)}^{(t-1)} = 0$, for example, if $i(t)$ has never been selected, then, by Moreau's identity, we also get a stochastic mirror descent step for $\frac{1}{2n}\sum_{i=1}^n d(x_i^\top \theta, \mathcal{Y}_i)^2$. However, this is not true anymore when an index is selected twice.

2.4 Accelerated coordinate descent

In the previous sections, we discussed randomized coordinate dual ascent to optimize the problem in Eq. (3). We can also consider accelerated proximal randomized coordinate ascent (Lin et al., 2015; Hendrikx et al., 2019; Allen-Zhu et al., 2016). For our problem, it leads to:

$$\mathbb{E}\Big[D_\psi(\theta^\star, \theta(\alpha^{(t)}))\Big] \leqslant \mathbb{E}\Big[\mathrm{gap}(\alpha^{(t)})\Big] \leqslant \frac{4\max_i L_i}{t^2}\left\{\frac{G(\alpha^\star) - G(0)}{\max_i L_i} + \frac{1}{2}\|\alpha^\star\|^2\right\}. \quad (8)$$

We will use the bound in Eq. (8) to analyze the general perceptron in the next section. We also provide the proximal accelerated randomized coordinate ascent algorithm (Lin et al., 2015; Hendrikx et al., 2019) with uniformly random sampling of coordinates to optimize the dual objective of ℓ_p-perceptron in Algorithm 2. However, the algorithm can easily be updated for the general case of Eqs. (3) and (4).

Note here that accelerated stochastic method for over-parametrized models in Algorithm 2 achieves Nesterov's fast rate without making explicit assumptions on the growth condition of the function and have the same computational overhead as that of primal SGD.

2.5 Baseline: Primal Mirror Descent

We will compare our dual algorithms to existing primal algorithms. They correspond to the minimization of

$$F(\theta) = \frac{1}{2n}\sum_{i=1}^n d(x_i^\top \theta, \mathcal{Y}_i)^2. \quad (9)$$

Vaswani et al. (2018) showed convergence of stochastic gradient descent for this problem. We extend their results to all mirror maps. Mirror descent with the mirror map ψ selects $i(t)$ at random and the iteration update is

$$\psi'(\theta^{(t)}) = \psi'(\theta^{(t-1)}) - \gamma x_{i(t)}(\Pi_{\mathcal{Y}_i}(x_{i(t)}^\top \theta^{(t-1)}) - x_{i(t)}^\top \theta^{(t-1)}). \quad (10)$$

Note that we have already encountered it in Lemma 2, for least-squares regression, where we provided a convergence rate on the final iterate.

In Theorem 1 below, we prove an $O(1/t)$ convergence rate for stochastic mirror descent update with mirror map ψ, for a constant step-size and the average iterate, directly extending the result of Vaswani et al. (2018) to all mirror maps.

Theorem 1 *Consider the stochastic mirror descent update in Eq. (10) for the optimization problem in Eq. (9) with $\gamma = \mu/\sup_i \|x_i\|_{2\to\star}^2$, the expected optimization error after t iterations the for averaged iterate $\bar{\theta}_t$ behaves as,*

$$0 \leqslant \mathbb{E}[F(\bar{\theta}^{(t)})] \leqslant \frac{\max_i L_i}{t}\psi(\theta^\star).$$

We provide the proof in Appendix A.2. The result is also applicable to general expectations and any form of convex objectives in the interpolation regime. We use this extension as one of our baseline in our experiments. In practice, as mentioned earlier, the update for mirror descent in Eq. (10) is similar to randomized dual coordinate ascent update in Lemma 1, in particular in early iterations (and not surprisingly, they behave similarly). Note here the difference in guarantees for the final iterates (which we get through a dual analysis) and the guarantees for the averaged iterate (which we get through a primal analysis).

3 ℓ_p-perceptrons

So far, we have discussed very general formulations for optimization problems in the interpolation regime. In this section, we discuss a specific problem which is widely used for linear binary classification, known as the perceptron algorithm, which is guaranteed to converge for linearly separable data. Here, we view

Algorithm 2 Accelerated Proximal Coordinate Ascent (Dual Perceptron) (Lin et al., 2015; Hendrikx et al., 2019)

Input: α_0, $\theta_0 \leftarrow \theta(\alpha_0)$, x_i for $i \in [n]$ and $\mu = 0$.
Initialize: $z_0 \leftarrow \alpha_0$, $\theta_{z_0} \leftarrow \theta_0$, $v_0 \leftarrow \alpha_0$ and $\gamma_0 \leftarrow \frac{1}{n}$.
Output: θ_{T+1} and α_{T+1}
for $t \leftarrow 0$ **to** T **do**
  Choose $i_t \in \{1, 2, \cdots, n\}$ randomly.
  $r_t = 1 - \theta_{z_t}^\top x_{i_t}$
  $\alpha_{t+1} = u_{t+1} = \alpha_t + \frac{r_t}{n\gamma_t L_{i_t}}$.
  $\alpha_{i(t)}^{(t+1)} = \max(\alpha_{i(t)}^{(t+1)}, 0)$.
  Update $\theta_{t+1} \leftarrow \theta(\alpha_{t+1})$. (Algorithm 3)
  $\gamma_{t+1} = \frac{1}{2}\left(\sqrt{\gamma_t^4 + 4\gamma_t^2} - \gamma_t^2\right)$.
  $v_{t+1} = z_t + n\gamma_t(\alpha_{t+1} - \alpha_t)$.
  $z_{t+1} = (1 - \gamma_{t+1})v_{t+1} + \gamma_{t+1}\alpha_{t+1}$.
  Update $\theta_{z_{t+1}} \leftarrow \theta(z_{t+1})$. (Algorithm 4)

Algorithm 3 Update $\boldsymbol{\theta}_{t+1}$

Input: x_{i_t},α_{t+1} , $X^\top \alpha_t$, α_t and i_t .
Output: θ_{t+1} and $X^\top \alpha_{t+1}$
  $X^\top \alpha_{t+1} = X^\top \alpha_t + (\alpha_{i(t)}^{(t+1)} - \alpha_{i(t)}^{(t)})x_{i_t}$.
  Compute θ_{t+1} from $X^\top \alpha_{t+1}$.

Algorithm 4 Update $\boldsymbol{\theta}_{\mathbf{z}_{t+1}}$

Input: x_{i_t}, α_{t+1}, $X^\top \alpha_t$, $X^\top \alpha_{t+1}$, $X^\top z_t$, α_t, γ_t, γ_{t+1} .
Output: $\boldsymbol{\theta}_{\mathbf{z}_{t+1}}$ and $X^\top z_{t+1}$
  $X^\top v_{t+1} = X^\top z_t + n\gamma_t X^\top(\alpha_{t+1} - \alpha_t)$.
  $X^\top z_{t+1} = (1 - \gamma_{t+1})X^\top v_{t+1} + \gamma_{t+1}X^\top \alpha_{t+1}$.
  Compute $\theta_{z_{t+1}}$ from $X^\top z_{t+1}$.

the generalized ℓ_p-norm perceptron algorithm from the lens of our primal-dual formulation.

We consider $(x_i, y_i) \in \mathbb{R}^d \times \{-1, 1\}$ for $i \in \{1, \cdots, n\}$, and the problem of minimizing $\psi(\theta)$ such that $\forall i, y_i x_i^\top \theta \geqslant 1$, which can be written as $\tilde{x}_i^\top \theta \geqslant 1$, where $\tilde{x}_i = y_i x_i$ for all $i \in \{1, \ldots, n\}$. For this section, we will be limiting ourselves to $\psi(\theta) = \frac{1}{2}\|\theta\|_p^2$ for $p \in (1, 2]$. We know that $\psi(\theta) = \frac{1}{2}\|\theta\|_p^2$ for $p \in (1, 2]$ is $(p-1)$-strongly convex with respect to the ℓ_p-norm. In this section, we denote $X \in \mathbb{R}^{n \times d}$ the data matrix $X = (\tilde{x}_1^\top; \tilde{x}_2^\top; \cdots; \tilde{x}_n^\top)$. Our generic optimization problem from Eq. (2) turns into:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|\theta\|_p^2 \text{ such that } X\theta \geqslant 1, \tag{11}$$

The dual problem is here

$$\max_{\alpha \in \mathbb{R}_+^n} -\frac{1}{2}\left\|\frac{-1}{n}\sum_{i=1}^n x_i \alpha_i\right\|_q^2 + \frac{1}{n}\sum_{i=1}^n \alpha_i, \tag{12}$$

where $\|\cdot\|_q$ is dual norm of $\|\cdot\|_p$, with $1/p + 1/q = 1$. At optimality, θ can be obtained from $X^\top \alpha$ as

$$\theta_j = \frac{1}{n}\|X^\top \alpha\|_q^{2-q}(X^\top \alpha)_j^{q-1},$$

where we define $u^{q-1} = |u|^{q-1}\text{sign}(u)$.

The function $\alpha \mapsto \frac{1}{2}\|X^\top \alpha\|_q^2$ is smooth, and the regular smoothness constant with respect to the i-th variable which is less than $L_i = \frac{1}{p-1}\|x_i\|_q^2$. We can apply here the results from Proposition 1 to get the convergence in primal iterates for the the general ℓ_p-norm perceptron formulation in Eq. (11), while optimizing the dual function via accelerated coordinate ascent in Eq. (12).

Corollary 1 *For the generalized ℓ_p-norm perceptron described in our primal-dual framework in Equations* (11) *and* (12)*, we have*

$$\mathbb{E}\Big[\|\theta(\alpha) - \theta^\star\|_p\Big] \le \sqrt{\frac{2\mathbb{E}[\text{gap}(\alpha)]}{p-1}}.$$

Proof The result comes from the application of Proposition 1 in the generalized ℓ_p-norm perceptron from setting Eq. (11), with $D_{\frac{1}{2}\|\cdot\|_p^2}(\theta^\star, \theta) \ge \frac{p-1}{2}\|\theta - \theta^\star\|_p^2$. ■

If we use accelerated randomized coordinate descent to optimize dual objective given in Eq. (12), then after t number of iterations, we get:

$$\mathbb{E}\Big[\|\theta_t - \theta^\star\|_p\Big] \le \frac{2\sqrt{2}\max_i \|x_i\|_q}{\sqrt{(p-1)}t}\sqrt{\frac{G(\alpha^\star) - G(0)}{\max_i \|x_i\|_q} + \frac{1}{2}\|\alpha^\star\|^2}, \tag{13}$$

where $\theta_t = \theta(\alpha_t)$.

Mistake bound. Since, we have the bound on the distance between primal iterate to its optimum, we can simply derive the mistake bound for our algorithm which we prove in Appendix B.

Lemma 3 *For the generalized ℓ_p-norm perceptron described in our primal-dual framework in Equations* (11) *and* (12)*, we make no mistakes on training data on average after*

$$t > \frac{2\sqrt{2}R^2}{\sqrt{p-1}}\sqrt{\frac{G(\alpha^\star) - G(0)}{R} + \frac{1}{2}\|\alpha^\star\|^2}$$

steps where $R = \max_i \|x_i\|_q$ and $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$.

The accelerated coordinate descent algorithm to solve the ℓ_p-perceptron is given in Algorithm 2. More details about the relationship between primal and dual variables, as well as dual ascent update for random coordinate descent for general ℓ_p-norm perceptron, e.g., the dual problem in Eq. (12), is given in Appendix B. Mistake bounds for the classical ℓ_p-perceptron are also recalled in Appendix B.

Baseline: primal mirror descent. We consider the finite sum minimization with stochastic mirror descent update and mirror map $\psi = \frac{1}{2}\|\cdot\|_p^2$ as discussed in Section 2.5, that is, the finite sum minimization in Eq. (9) with $f_i(\theta) = \frac{1}{2}(1-\theta^\top x_i)_+^2$.

Corollary 2 *Consider the finite sum minimization of* $f(\theta) = \frac{1}{2n}\sum_{i=1}^n (1-\theta^\top x_i)_+^2$ *via stochastic mirror descent with mirror map* $\psi(\cdot) = \frac{1}{2}\|\cdot\|_p^2$, *then on average, the proportion of mistakes on the training set is less than* $\sqrt{\frac{\|\theta^\star\|_p^2 R^2}{(p-1)t}}$ *where* $R = \max_i \|x_i\|_q$.

Proof The proof comes directly from Theorem 1 and from the fact that the proportion of mistakes on the training set is less than the square root of the excess risk. ∎

Similar bounds on the proportion of mistakes can also be obtained while optimizing $f(\theta) = \frac{1}{n}\sum_{i=1}^n (1-x_i^\top\theta)_+$ via stochastic mirror descent with mirror map $\frac{1}{2}\|\cdot\|_p^2$. However, while tuning the step size, it requires the knowledge of $\|\theta^\star\|_p$, hence we do not include it in our base line.

We can compare the minimum number of iterations required to achieve no further mistakes while training in Lemma 3 and Corollary 2 to get the conditions on optimal primal and dual optimal variables under which our method (which has a better dependence in the number of iterations t) performs better than the baseline. We discuss these in the Appendix B. In our empirical evaluationin Section 4, dual accelerate coordinate ascent significantly outperforms primal mirror descent.

Special Case of ℓ_1-perceptron. Our goal in this specific case is to solve the following sparse problem,

$$\theta_0 = \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2}\|\theta\|_1^2 \text{ such that } X\theta \geqslant 1. \qquad (14)$$

$\|\cdot\|_1$ is not strongly convex, hence we can not fit this problem to our formulation. However, following Duchi et al. (2010); Ball et al. (1994), we solve the problem in (11) with $p = 1 + \frac{1}{\log d}$ where d is the dimension.

4 Experiments

In this section, we provide empirical evaluation for the methods discussed in this paper with the ℓ_p-perceptron. We generate data from a Gaussian distribution in dimension $d = 2000$, which we describe below. We consider two settings of p for our experiments, $p = 2$ which is usual perceptron, and $p = 1 + \frac{1}{\log d}$, which is the sparse perceptron setting.

Data generation. We generate $n = 1000$ inputs $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$ with $d = 2000$ dimensional from a Gaussian distribution centered at 0 and covariance matrix Σ which is a diagonal matrix. Similarly, we generate a random $d = 2000$ prediction vector θ sampled again from the normal distribution.

For ℓ_2-perceptron, the i-th eigenvalue for Σ is $1/i^{3/2}$ and for sparse perceptron i-th eigenvalue for Σ, is $1/i$. We compute the prediction vector y_i for x_i as follows, $y_i = \text{sign}(x_i^\top\theta + b)$ where we fix $b = 0.005$. We also remove those pair of (x_i, y_i) from the data for which we have $x_i^\top\theta + b \leq 0.1$. We generate 1000 train examples and 1000 test examples for both settings. For the sparse perceptron case, we make the prediction vector θ sparse by randomly choosing 50 entries to be non zero. We then compute the prediction vector similar to the ℓ_p-perceptron case, $y_i = \text{sign}(x_i^\top\theta + b)$ where we fix $b = 0.005$ and remove those pair of (x_i, y_i) from the data for which we have, $x_i^\top\theta + b \leq 0.1$.

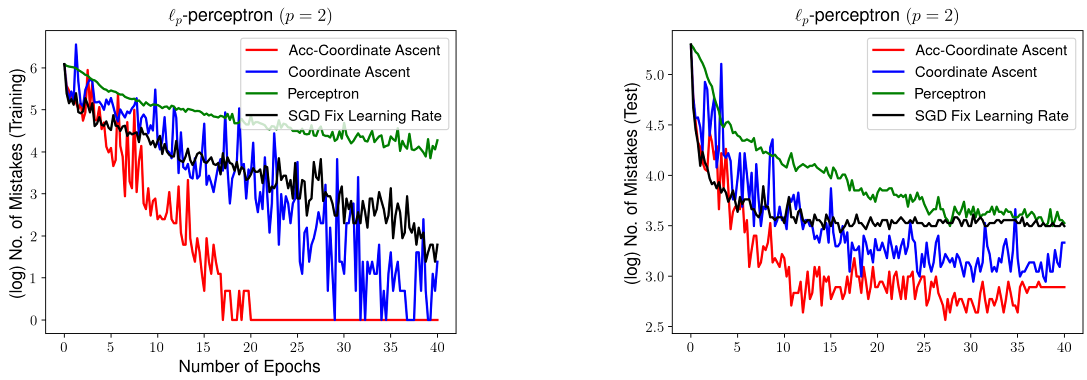
Baseline. For the ℓ_2-perceptron, we compare accelerated coordinate descent and randomized coordinate descent with the perceptron and primal SGD (Vaswani et al., 2018). For the sparse perceptron, we compare the accelerated coordinate descent and randomized coordinate descent with extension of primal SGD to stochastic mirror descent case (discussed in section 2.5 with $f_i = \frac{1}{2}(1-x_i^\top\theta)_+^2$) with mirror map $\psi(\cdot) = \frac{1}{2}\|\cdot\|_p^2$ where $p = 1 + \frac{1}{\log d}$. Note that we compare to non-averaged SGD (for which we provide a new proof), which works significantly better than averaged SGD.

Comparisons for the ℓ_2-perceptron and sparse perceptron are given in Figures 1 and 2 respectively.

We can make the following observations:

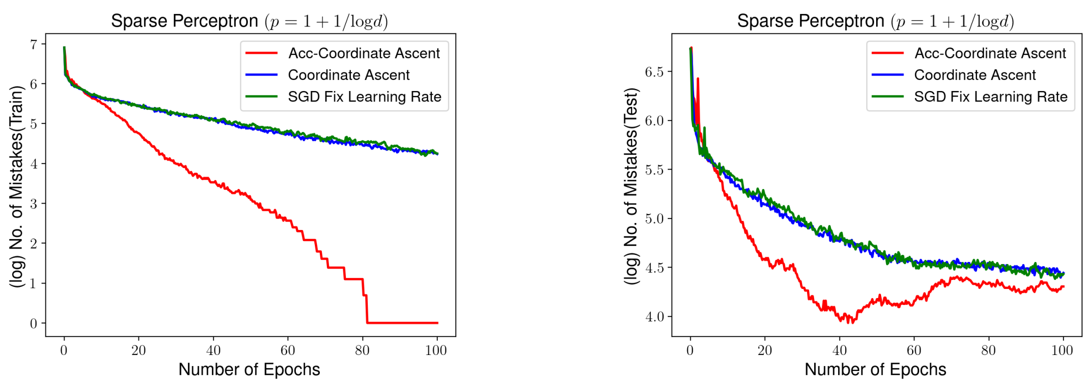
(a) From both the training plots (Figure 1a and Figure 2a), it is clear that we gain significantly in training performance over primal SGD and the perceptron if we optimize the dual with accelerated randomized coordinate ascent method, which supports our theoretical claims made in Section 3.

(b) For testing errors, we also see gains for our accelerated perceptron, which is not supported by



(a) Number of mistakes on the training test (in log scale). (b) Number of mistakes on the test (in log scale).

Figure 1: Experimental results for ℓ_2-perceptron



(a) Number of mistakes on the training (in log scale). (b) Number of mistakes on the test (in log scale).

Figure 2: Experimental results for sparse perceptron.

theoretical arguments. This gives motivation to further study this algorithm for general expectations.

(c) Note that in the semi-log plots, we observe an affine behavior of the training errors, highlighting exponential convergence. This can be explained by a strongly convex dual problem (since the matrix $XX^\top$ is invertible), and could be quantified using usual convergence rates for coordinate ascent for strongly-convex objectives.

5 Conclusion

In this paper, we proposed algorithms that are explicitly regularizing solutions of an interpolation problem. This is done through a dual approach, and, with acceleration, it improves over existing algorithms. Several natural questions are worth exploring: (1) Can we explicitly characterize linear convergence in the dual (like observed in experiments), with or without regularization? (2) How are our algorithms performing beyond the interpolation regime, where the dual become unbounded but some primal information can typically be recovered in Dykstra-style algorithms (Bauschke & Koch, 2015)? (3) Can we extend our approach to saddle-point formulations such as proposed by Kundu et al. (2018)? Can we prove any improvement in the general population regime, where we aim at bounds on testing data?

In a recent series of papers (Ergen & Pilanci, 2020; Pilanci & Ergen, 2020), it was shown that strong duality holds for different neural network architecture and there might be a possibility to extend our approach for those architectures which are beyond the convexity assumptions. It is a promising future research direction and needs to be investigated further.

Acknowledgements

This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA

Institute). We also acknowledge support from the European Research Council (grant SEQUOIA 724063).

References

Allen-Zhu, Z., Qu, Z., Richtárik, P., and Yuan, Y. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pp. 1110–1119, 2016.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 7411–7422, 2019.

Bach, F. and Moulines, E. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pp. 451–459, 2011.

Ball, K., Carlen, E. A., and Lieb, E. H. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994.

Bassily, R., Belkin, M., and Ma, S. On exponential convergence of sgd in non-convex over-parametrized learning. *arXiv preprint arXiv:1811.02564*, 2018.

Bauschke, H. H. and Combettes, P. L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

Bauschke, H. H. and Koch, V. R. Projection methods: Swiss army knives for solving feasibility and best approximation problems with halfspaces. *Contemp. Math*, 636:1–40, 2015.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*, 2018.

Block, H.-D. The perceptron: A model for brain functioning. i. *Reviews of Modern Physics*, 34(1):123, 1962.

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

Boyle, J. P. and Dykstra, R. L. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In *Advances in Order Restricted Statistical Inference*, pp. 28–47. Springer, 1986.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Calatroni, L., Garrigos, G., Rosasco, L., and Villa, S. Accelerated iterative regularization via dual diagonal descent. *arXiv preprint arXiv:1912.12153*, 2019.

Cevher, V. and Vũ, B. C. On the linear convergence of the stochastic gradient method with constant step-size. *Optimization Letters*, 13(5):1177–1187, 2019.

Chambolle, A., Tan, P., and Vaiter, S. Accelerated alternating descent methods for Dykstra-like problems. *Journal of Mathematical Imaging and Vision*, 59(3): 481–497, 2017.

Defazio, A., Bach, F., and Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Duchi, J. C., Shalev-Shwartz, S., Singer, Y., and Tewari, A. Composite objective mirror descent. In *COLT*, pp. 14–26, 2010.

Ergen, T. and Pilanci, M. Training convolutional relu neural networks in polynomial time: Exact convex optimization formulations. *arXiv preprint arXiv:2006.14798*, 2020.

Flammarion, N. and Bach, F. Stochastic composite least-squares regression with convergence rate o (1/n). *arXiv preprint arXiv:1702.06429*, 2017.

Freund, Y. and Schapire, R. E. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296, 1999.

Gaffke, N. and Mathar, R. A cyclic projection algorithm via duality. *Metrika*, 36(1):29–54, 1989.

Golub, G. H., Hansen, P. C., and O'Leary, D. P. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1):185–194, 1999.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Grove, A. J., Littlestone, N., and Schuurmans, D. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, 2018.

Halperin, I. The product of projection operators. *Acta Sci. Math.(Szeged)*, 23(1):96–99, 1962.

Hendrikx, H., Bach, F., and Massoulié, L. An accelerated decentralized stochastic proximal algorithm for finite sums. In *Advances in Neural Information Processing Systems*, pp. 952–962, 2019.

Kakade, S., Shalev-Shwartz, S., and Tewari, A. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2(1), 2009.

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

Kivinen, J. Online learning of linear classifiers. In *Advanced lectures on machine learning*, pp. 235–257. Springer, 2003.

Kubo, M., Banno, R., Manabe, H., and Minoji, M. Implicit regularization in over-parameterized neural networks. *arXiv preprint arXiv:1903.01997*, 2019.

Kundu, A., Bach, F., and Bhattacharya, C. Convex optimization over intersection of simple sets: improved convergence rate guarantees via an exact penalty approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 958–967, 2018.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

Lin, Q., Lu, Z., and Xiao, L. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4): 2244–2273, 2015.

Liu, C. and Belkin, M. Accelerating SGD with momentum for over-parameterized learning. *arXiv preprint arXiv:1810.13395*, 2018.

Ma, S., Bassily, R., and Belkin, M. The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning. *arXiv preprint arXiv:1712.06559*, 2017.

Minsky, M. and Papert, S. A. *Perceptrons: An introduction to computational geometry*. MIT press, 2017.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Nesterov, Y. and Stich, S. U. Efficiency of the accelerated coordinate descent method on structured optimization problems. *SIAM Journal on Optimization*, 27(1):110–123, 2017.

Pilanci, M. and Ergen, T. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. *arXiv preprint arXiv:2002.10553*, 2020.

Polyak, B. T. New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2, 1990.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM Fournal on Control and Pptimization*, 30(4):838–855, 1992.

Richtárik, P. and Takáč, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Rudi, A., Camoriano, R., and Rosasco, L. Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pp. 1657–1665, 2015.

Shalev-Shwartz, S. and Singer, Y. A new perspective on an old perceptron algorithm. In *International Conference on Computational Learning Theory*, pp. 264–278. Springer, 2005.

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

Shalev-Shwartz, S. and Zhang, T. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International conference on machine learning*, pp. 64–72, 2014.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Tibshirani, R. J. Dykstra's algorithm, admm, and coordinate descent: Connections, insights, and extensions. In *Advances in Neural Information Processing Systems*, pp. 517–528, 2017.

Tsampouka, P. and Shawe-Taylor, J. Analysis of generic perceptron-like large margin classifiers. In *European Conference on Machine Learning*, pp. 750–758. Springer, 2005.

Tseng, P. Dual coordinate ascent methods for non-strictly convex minimization. *Mathematical programming*, 59(1-3):231–247, 1993.

Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

Tseng, P. and Bertsekas, D. P. Relaxation methods for problems with strictly convex separable costs and linear constraints. *Mathematical Programming*, 38 (3):303–321, 1987.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G., and Lacoste-Julien, S. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *arXiv preprint arXiv:1905.09997*, 2019.

Von Neumman, J. Functional operators ii, the geometry of orthogonal spaces. *Annals of Math. studies*, 22, 1951.

Ward, R., Wu, X., and Bottou, L. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pp. 6677–6686, 2019.

Weese, J. A regularization method for nonlinear ill-posed problems. *Computer Physics Communications*, 77(3):429–440, 1993.

Yao, Y., Rosasco, L., and Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Zhang, J., Rivard, B., and Rogge, D. The successive projection algorithm (SPA), an algorithm with a spatial constraint for the automatic search of endmembers in hyperspectral data. *Sensors*, 8(2):1321–1342, 2008.