Implicit Regularization Properties of Variance Reduced Stochastic Mirror Descent

Yiling Luo, Xiaoming Huo, and Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology
{yluo373, huo, ymei3}@gatech.edu

***Abstract*—In machine learning and statistical data analysis, we often run into objective function that is a summation: the number of terms in the summation possibly is equal to the sample size, which can be enormous. In such a setting, the stochastic mirror descent (SMD) algorithm is a numerically efficient method—each iteration involving a very small subset of the data. The variance reduction version of SMD (VRSMD) can further improve SMD by inducing faster convergence. On the other hand, algorithms such as gradient descent and stochastic gradient descent have the implicit regularization property that leads to better performance in terms of the generalization errors. Little is known on whether such a property holds for VRSMD. We prove here that the discrete VRSMD estimator sequence converges to the minimum mirror interpolant in the linear regression. This establishes the implicit regularization property for VRSMD. As an application of the above result, we derive a model estimation accuracy result in the setting when the true model is sparse. We use numerical examples to illustrate the empirical power of VRSMD.**

***Index Terms*—implicit regularization, variance reduction, stochastic mirror descent, overparameterization.**

I. Introduction

In statistics and machine learning, it is common to optimize an objective function that is a finite-sum. SMD efficiently optimizes such an objective by using a subset of data to do one step update of the variable/parameter. Further adopting the variance reduction technique to SMD, we get the VRSMD algorithm that enjoys fast convergence [13, 3].

The implicit regularization is a relatively new concept [7] that explains why a result of an algorithm generalizes well in some overparameterized models [7, 9]. It refers to the fact that an algorithm can automatically select a minimum norm solution, which is not explicitly induced by the objective function. There are works on implicit regularization for Gradient Descent [18, 21, 10, 4], Stochastic Gradient Descent [1, 6, 17, 8], and SMD [5]. Given the computational advantage of VRSMD compared to the algorithms above, it would be better if VRSMD also has the useful implicit regularization property.

From technical point of view, our paper contains two results:

- In linear regression (including underfitting and overfitting), we show that the solution sequence of VRSMD converges to the minimum mirror interpolant, which is the implicit regularization property of VRSMD, and we also specify the convergence rate.
- In sparse regression, by choosing a proper mirror map, we show that the implicit regularization estimator finds the sparse true parameter with a small error. Moreover, compared with the deterministic algorithms in [18, 21], our algorithm is equally good in estimating a sparse truth while being computationally faster, as supported by our experiments.

From the application point of view, our result shows that the Mirror Descent and its variants are useful to explore the low dimensional geometric structure from high dimensional data, which leads to nice generalization properties.

Notation. The following notations are used throughout this paper. For a matrix $X \in \mathbb{R}^{n\times p}$, we denote by $\mathrm{col}(X) := \{\mathbf{u} \in \mathbb{R}^n : \exists \mathbf{v} \in \mathbb{R}^p, \mathbf{u} = X\mathbf{v}\}$ the column space of X, and we denote by $\mathcal{N}(X) := \{\mathbf{v} \in \mathbb{R}^p : X\mathbf{v} = \mathbf{0}\}$ the null space of X. For a vector $\mathbf{v} \in \mathbb{R}^p$, we use the definition of ℓ_p norm of $\mathbf{v}$ that $\|\mathbf{v}\|_p = (\sum_i |v_i|^p)^{1/p}$ for $p \geq 1$ and we denote the number of non-zero elements in $\mathbf{v}$ as $\|\mathbf{v}\|_0$. For a subset of indexes $I \subset \{1, \ldots, p\}$, we define $\mathbf{v}_I := (v_i)_{i\in I}$, and denote the cardinality of I as $|I|$. For a set $\mathcal{X} \subset \mathbb{R}^p$, we define $P_{\mathcal{X}}\mathbf{v} = \mathrm{argmin}_{\mathbf{u}\in\mathcal{X}} \|\mathbf{u} - \mathbf{v}\|_2$. For two non-negative-valued functions $a(x)$ and $b(x)$, we denote $a(x) \sim \mathcal{O}(b(x))$ if there exists an absolute constant C such that $a(x) \leq Cb(x)$; and we denote $a(x) \sim \Theta(b(x))$ if there are absolute constants c, C such that $cb(x) \leq a \leq Cb(x)$.

Paper Organization. The rest of the paper is organized as follows. In Section II, we describe our problem formulation and algorithm. Section III states the main theory on the implicit regularization. Section IV develops insight into the implicit regularization and establishes the sparse recovery property. Section V supports the theory on implicit regularization by simulations and experiments. In Section VI, we discuss the finding of our work and some future directions. Due to page limit, we only include necessary description of experiment and proof sketch. Full proofs and experiment details can be found in our arXiv paper [14].

II. Formulation and Algorithm

To better present the material, we split this section into two subsections. In Subsection II-A we formulate the optimization

This project is partially supported by the Transdisciplinary Research Institute for Advancing Data Science (TRIAD), http://triad.gatech.edu, which is a part of the TRIPODS program at NSF and locates at Georgia Tech, enabled by the NSF grant CCF-1740776. Luo is supported in part by ARC fellowship. Huo is supported in part by NSF grant DMS-2015363. Mei is supported in part by NSF grant DMS-2015405.

problem motivated from linear regression. In Subsection II-B, we present the Variance Reduction Stochastic Mirror Descent (VRSMD) algorithm for solving such an optimization problem.

A. Formulation

Assume we observe data pairs $\{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}_{i=1}^n$, the goal is to predict the response y based on $\mathbf{x}$. Under the empirical risk minimization framework, we consider the general optimization problem of the form

$$\min_{\boldsymbol{\beta}} F(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{\beta}; (\mathbf{x}_i, y_i)), \tag{1}$$

and we shorten $f_i(\boldsymbol{\beta}; (\mathbf{x}_i, y_i))$ as $f_i(\boldsymbol{\beta})$ to simplify the notation.

As a concrete example, for the linear regression model, the classical least squares method is to find coefficient $\boldsymbol{\beta}$ that minimizes the objective function

$$\min_{\boldsymbol{\beta}} F(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}\left(\mathbf{x}_i^T\boldsymbol{\beta} - y_i\right)^2 = \frac{1}{2n}\|X\boldsymbol{\beta} - y\|_2^2, \tag{2}$$

where we denote $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times p}$, and $\mathbf{y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$.

When problem (2) has non-unique solutions, it is important yet nontrivial to find a solution that has nice generalization property. It is well known that when running Gradient Descent algorithm with initialization $\boldsymbol{\beta}_0 = \mathbf{0}$ on (2), the corresponding solution is the minimal ℓ_2 norm solution among all solutions of (2), see [20]. Also, the properties of SMD are studied in [5]. It is unknown what happens to the solution if we run variants of SMD, for example, variance reduced SMD.

B. VRSMD Algorithm

Let us now present the main idea of the variance reduced stochastic mirror descent (VRSMD) algorithm as follows.

To understand why we need variance reduction, consider the Stochastic Mirror Descent(SMD) algorithm using a strictly convex and differentiable mirror map $\psi(\cdot)$. At step t, the SMD updates $\boldsymbol{\beta}_{t+1}$ such that

$$\nabla\psi(\boldsymbol{\beta}_{t+1}) = \nabla\psi(\boldsymbol{\beta}_t) - \eta_t \nabla f_{i_t}(\boldsymbol{\beta}_t),$$

where i_t is randomly sampled from $\{1, \ldots, n\}$. Now the term $\nabla f_{i_t}(\boldsymbol{\beta}_t)$ has $\mathbb{E}[\nabla f_{i_t}(\boldsymbol{\beta}_t)] = \nabla F(\boldsymbol{\beta}_t)$, so SMD has unbiased update compared to Mirror Descent, where the update is $\nabla\psi(\boldsymbol{\beta}_{t+1}) = \nabla\psi(\boldsymbol{\beta}_t) - \eta_t\nabla F(\boldsymbol{\beta}_t)$. However, in general $\mathrm{Var}[\nabla f_{i_t}(\boldsymbol{\beta}_t)] \neq 0$ for any $\boldsymbol{\beta}_t$, so we need $\eta_t \to 0$, which may lead to slow convergence.

Variance reduction addresses the issues above by replacing $\nabla f_{i_t}(\boldsymbol{\beta}_t)$ with term A_t such that

- $\mathbb{E}[A_t] = \mathbb{E}[\nabla f_{i_t}(\boldsymbol{\beta}_t)]$ to keep unbiased update;
- $\mathrm{Var}[A_t] < \mathrm{Var}[\nabla f_{i_t}(\boldsymbol{\beta}_t)]$ to control variance.

One choice of A_t is $A_t = \nabla f_{i_t}(\boldsymbol{\beta}_t) - B_t + \mathbb{E}[B_t]$, where B_t and $\nabla f_{i_t}(\boldsymbol{\beta}_t)$ are positively correlated with correlation coefficient $r > 0.5$ and $\mathrm{Var}[B_t] \approx \mathrm{Var}[\nabla f_{i_t}(\boldsymbol{\beta}_t)]$. For this A_t one can check that $\mathbb{E}[A_t] = \mathbb{E}[\nabla f_{i_t}(\boldsymbol{\beta}_t)]$ and $\mathrm{Var}[A_t] = \mathrm{Var}[\nabla f_{i_t}(\boldsymbol{\beta}_t) - B_t] = \mathrm{Var}[\nabla f_{i_t}(\boldsymbol{\beta}_t)] - 2r\sqrt{\mathrm{Var}[\nabla f_{i_t}(\boldsymbol{\beta}_t)]\,\mathrm{Var}[B_t]} + \mathrm{Var}[B_t] < \mathrm{Var}[\nabla f_{i_t}(\boldsymbol{\beta}_t)]$. For a proper B_t such that $\mathrm{Var}(A_t) \overset{t\to\infty}{\longrightarrow} 0$, the algorithm converges for a fixed η.

For illustration purpose, we use the variance reduction technique in [12] to get the VRSMD Algorithm 1. However, one should note that this framework applies to other variance reduction methods such as SARAH [15] and SPIDER [11].

Algorithm 1: Variance Reduced Stochastic Mirror Descent (VRSMD)

Input: An objective function $F(\cdot) = \frac{1}{n}\sum_{i=1}^n f_i(\cdot)$, and a strictly convex and differentiable mirror map $\psi(\cdot)$;

Initialization: Initialize $\tilde{\boldsymbol{\beta}}^0$. Choose the step-size η, outer iteration number S, inner iteration number m. Denote the estimator at tth inner iteration of sth outer iteration as $\boldsymbol{\beta}_t^s$. Set $\boldsymbol{\beta}_1^1 = \boldsymbol{\beta}_{m+1}^0 = \tilde{\boldsymbol{\beta}}^0$;

for *Outer iteration s = 1,…,S* **do**

  Calculate $\nabla F(\tilde{\boldsymbol{\beta}}^{s-1})$;

  for *Inner iteration t = 1,…,m* **do**

    Randomly sample i_t from $\{1, \ldots, n\}$, calculate

$$\mathbf{v}_t^s = \nabla f_{i_t}(\boldsymbol{\beta}_t^s) - \nabla f_{i_t}(\tilde{\boldsymbol{\beta}}^{s-1}) + \nabla F(\tilde{\boldsymbol{\beta}}^{s-1}), \tag{3}$$

    and update $\boldsymbol{\beta}_{t+1}^s$ such that

$$\nabla\psi(\boldsymbol{\beta}_{t+1}^s) = \nabla\psi(\boldsymbol{\beta}_t^s) - \eta\mathbf{v}_t^s. \tag{4}$$

  Set $\tilde{\boldsymbol{\beta}}^s$ to be a uniform random sample from $\{\boldsymbol{\beta}_1^s, \ldots, \boldsymbol{\beta}_m^s\}$;

  Option I: Set $\boldsymbol{\beta}_1^{s+1} = \boldsymbol{\beta}_{m+1}^s$;

  Option II: Set $\boldsymbol{\beta}_1^{s+1} = \tilde{\boldsymbol{\beta}}^s$;

Option I: Output $\boldsymbol{\beta}_a$ chosen uniformly random from $\{\{\boldsymbol{\beta}_t^s\}_{t=1}^m\}_{s=1}^S$;

Option II: Output $\boldsymbol{\beta}_a = \tilde{\boldsymbol{\beta}}^S$.

Remark 1. *We note the complexity of VRSMD as follows: The total number of stochastic-first-order calls (i.e. SFO complexity) of Algorithm 1 is $\mathcal{O}(nS + mS)$. A popular choice of the inner loop number m is $\Theta(n)$, which leads to $\mathcal{O}(1)$ SFO complexity per inner loop.*

The variance reduction component of VRSMD is $\mathbf{v}_t^s$ in (3). Note that it is equivalent to the variance reduction scheme we discussed above by taking $B_t = \nabla f_{i_t}(\tilde{\boldsymbol{\beta}}^{s-1})$. The conditions we list there on B_t hold when $\boldsymbol{\beta}_t^s$ and $\tilde{\boldsymbol{\beta}}^{s-1}$ are close, which happens by taking moderate values of the inner iteration number m and the step-size η.

We show that the VRSMD algorithm is a generalization of the SVRG algorithm in [12, 16]: For the special case of $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, we have $\nabla\psi(\boldsymbol{\beta}) = \boldsymbol{\beta}$, then (3) updates $\boldsymbol{\beta}_{t+1}^s$ as:

$$\boldsymbol{\beta}_{t+1}^s = \boldsymbol{\beta}_t^s - \eta\mathbf{v}_t^s,$$

which is the SVRG update, and VRSMD reduces to SVRG.

III. Implicit Regularization

In this section, we present the implicit regularization property of the VRSMD solution. To do so, it is necessary to first show that the VRSMD converges.

To begin with, we introduce some definitions that will be useful in our theory.

Definition 1 (L-smoothness)**.** *f is L-smooth with respect to $\|\cdot\|$ norm if there exists a constant $L>0$ such that*

$$\|\nabla f(\mathbf{u})-\nabla f(\mathbf{w})\|_* \le L\|\mathbf{u}-\mathbf{w}\|, \forall \mathbf{u},\mathbf{w},$$

where $\|\cdot\|_* := \max_{y:\|y\|=1}\langle y,\cdot\rangle$ is the dual norm of $\|\cdot\|$.

Definition 2 (α-strongly convex)**.** *f is α-strongly convex with respect to $\|\cdot\|$ norm if there exists a constant $\alpha>0$ such that*

$$f(\mathbf{u}) \ge f(\mathbf{w}) + \nabla f(\mathbf{w})^T(\mathbf{u}-\mathbf{w}) + \frac{\alpha}{2}\|\mathbf{u}-\mathbf{w}\|^2, \forall \mathbf{u},\mathbf{w}.$$

Definition 3 (Quadratic growth (QG))**.** *Let $\mathcal{X}$ be the set of all minimizers of f. f satisfies QG condition w.r.t. $\|\cdot\|$ if*

$$\frac{\mu}{2}\|\mathbf{u}-P_{\mathcal{X}}\mathbf{u}\|^2 \le f(\mathbf{u})-f(P_{\mathcal{X}}\mathbf{u}), \forall \mathbf{u}. \tag{5}$$

Definition 4 (ϵ-solution)**.** *For the optimization problem*

$$Opt = \min_{x\in B}\{f(x): g_i(x)\le 0, 1\le i\le m\}. \tag{6}$$

$x_\epsilon \in B$ is called an ϵ-solution to (6) *if*

$$f(x_\epsilon)-Opt \le \epsilon,$$
$$g_i(x_\epsilon)\le \epsilon, 1\le i\le m.$$

Definition 5 (Restricted eigenvalue (RE))**.** *X satisfies (s,γ)-RE condition if for any $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}\|_0 \le s$ we have*

$$\frac{\frac{1}{n}\|X\boldsymbol{\beta}\|_2^2}{\|\boldsymbol{\beta}\|_2^2} \ge \gamma.$$

Definition 6 (s-good)**.** *A matrix $X_{n\times p}$ is s-good if $\exists \kappa < \frac{1}{2}$ is such that $\forall \mathbf{u}\in\mathcal{N}(X)\subset\mathbb{R}^p$ and $\forall I\subset\{1,\dots,p\}$ with $|I|\le s$, we have*

$$\|\mathbf{u}_I\|_1 \le \kappa\|\mathbf{u}\|_1.$$

Next, let us present the convergence result of VRSMD:

Proposition 1. *Assume $F(\cdot)=\frac{1}{n}\sum_{i=1}^n f_i(\cdot)$ has every $f_i(\cdot)$ convex and L-smooth w.r.t. an arbitrary norm $\|\cdot\|$, and $\psi(\cdot)$ is α-strongly convex w.r.t. $\|\cdot\|$. Denote $\boldsymbol{\beta}^*=\arg\min F(\cdot)$.*

***(a)** Run Option I of Algorithm 1 on F with $\eta<\frac{\alpha}{24L}$, then we have*

$$\mathbb{E}[F(\boldsymbol{\beta}_a)-F(\boldsymbol{\beta}^*)] \le \frac{\alpha}{(\alpha\eta-24L\eta^2)T}\times \left[D_\psi(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\beta}}^0)+\frac{12L\eta^2 m}{\alpha}(F(\tilde{\boldsymbol{\beta}}^0)-F(\boldsymbol{\beta}^*))\right], \tag{7}$$

where $T=m\cdot S$, and $D_\psi(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\beta}}^0) := \psi(\boldsymbol{\beta}^*)-\psi(\tilde{\boldsymbol{\beta}}^0)-\langle\nabla\psi(\tilde{\boldsymbol{\beta}}^0),\boldsymbol{\beta}^*-\tilde{\boldsymbol{\beta}}^0\rangle$ is the Bregman divergence.

***(b)** If we further assume that $F(\cdot)$ satisfies the QG condition in* (5) *with constant μ, and that $\psi(\cdot)$ is ℓ-smooth, all w.r.t. $\|\cdot\|$, and also suppose that we run Option II of Algorithm 1 with a large enough m such that*

$$\tau := \frac{12L\eta^2/\alpha+\ell/(m\mu)}{\eta-12L\eta^2/\alpha} < 1, \tag{8}$$

then the VRSMD has a stronger linear convergence rate:

$$\mathbb{E}[F(\boldsymbol{\beta}_a)-F(\boldsymbol{\beta}^*)] \le \tau^S[F(\tilde{\boldsymbol{\beta}}^0)-F(\boldsymbol{\beta}^*)]. \tag{9}$$

Remark 2. *We analyze the computational complexity implied by Proposition 1: In (a), let $m=n$ and take $\eta=\frac{\alpha}{48L}$, then we have*

$$\mathbb{E}[F(\boldsymbol{\beta}_a)-F(\boldsymbol{\beta}^*)] \le \frac{96L}{\alpha T}\times \left[D_\psi(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\beta}}^0)+\frac{\alpha n}{192L}(F(\tilde{\boldsymbol{\beta}}^0)-F(\boldsymbol{\beta}^*))\right].$$

In this case, the number of gradient computations for achieving an ϵ-solution is $\mathcal{O}(\frac{L}{\epsilon}+\frac{n}{\epsilon})$.

Remark 3. *The assumption in (b) is moderate. Take $m=\frac{110L\ell}{\alpha\mu}$ and $\eta=\frac{\alpha}{36L}$, we have $\tau<1$. This choice of m does not violate the $\mathcal{O}(1)$ SFO comlexity per iteration – take a good mirror map so that $\ell/\alpha=\mathcal{O}(1)$, and consider the most indicative case [12] where the condition number $L/\mu=n$, we have $m=\Theta(n)$.*

Our results in Proposition 1 are consistent with those for SVRG. In part (a), the $\mathcal{O}(1/T)$ convergence rate matches the rate in [16]; in part (b), the linear rate matches the rate in [12], while we reduce their strong convexity assumption to quadratic growth.

Finally, we are ready to present our main result on implicit regularization. In the following theorem, we show that VRSMD finds an ϵ-solution of the minimum mirror interpolation problem.

Theorem 1. *For the objective function in* (2)*, assume $\psi(\boldsymbol{\beta})$ is α-strongly convex w.r.t. $\|\cdot\|_2$, denote $L=\max_i\|\mathbf{x}_i\|_2^2$ and let s_m be the smallest nonzero singular value of X. The VRSMD algorithm converges to the minimum mirror map interpolant*

$$\boldsymbol{\beta}^\psi := \operatorname{argmin}_{\boldsymbol{\beta}} \quad \psi(\boldsymbol{\beta}) \quad s.t. \quad F(\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}'} F(\boldsymbol{\beta}'). \tag{10}$$

We describe the convergence by the following ϵ-solution:

***(a)** Run Option I of Algorithm 1 with choice $\eta<\frac{\alpha}{24L}$ and initialization $\tilde{\boldsymbol{\beta}}^0$ such that $\nabla\psi(\tilde{\boldsymbol{\beta}}^0)\in\mathrm{col}(X^T)$, assume that the output $\boldsymbol{\beta}^a$ satisfies $\|\nabla\psi(\boldsymbol{\beta}^a)\|_2\le B$, then we will have*

$$\mathbb{E}[\psi(\boldsymbol{\beta}^a)-\psi(\boldsymbol{\beta}^\psi)] \le \frac{B}{s_m}\sqrt{\frac{\alpha}{(\alpha\eta-24L\eta^2)T}}\times \left[2nD_\psi(\boldsymbol{\beta}^\psi,\tilde{\boldsymbol{\beta}}^0)+\frac{24nL\eta^2 m}{\alpha}\left(F(\tilde{\boldsymbol{\beta}}^0)-F(\boldsymbol{\beta}^\psi)\right)\right]^{.5}, \tag{11}$$

which describes how far is the objective value at $\boldsymbol{\beta}^a$ away from the optimal solution to (10)*. Moreover, we have*

$$\mathbb{E}[F(\boldsymbol{\beta}^a)-F(\boldsymbol{\beta}^\psi)] \le \frac{\alpha}{(\alpha\eta-8L\eta^2)T}\times \left[D_\psi(\boldsymbol{\beta}^\psi,\tilde{\boldsymbol{\beta}}^0)+\frac{12L\eta^2 m}{\alpha}\left(F(\tilde{\boldsymbol{\beta}}^0)-F(\boldsymbol{\beta}^\psi)\right)\right], \tag{12}$$

which characterizes how much does $\boldsymbol{\beta}^a$ violates the constraints of (10)*. They together show that the VRSMD algorithm finds an ϵ-solution to* (10) *for $T = \mathcal{O}(\frac{1}{\epsilon} + \frac{1}{\epsilon^2})$.*

***(b)** Further assume that $\psi(\cdot)$ is ℓ-smooth w.r.t. $\|\cdot\|_2$ and*

$$\tau' = \frac{12L\eta^2/\alpha + \ell n/(ms_m^2)}{\eta - 12L\eta^2/\alpha} < 1. \tag{13}$$

Run Option II of Algorithm 1, we can show that:

$$\mathbb{E}[\psi(\boldsymbol{\beta}^a) - \psi(\boldsymbol{\beta}^\psi)] \le \frac{B(\tau')^{S/2}\sqrt{2n}}{s_m}\sqrt{F(\tilde{\boldsymbol{\beta}}^0) - F(\boldsymbol{\beta}^\psi)},$$
$$\mathbb{E}[F(\boldsymbol{\beta}^a) - F(\boldsymbol{\beta}^\psi)] \le (\tau')^S \left(F(\tilde{\boldsymbol{\beta}}^0) - F(\boldsymbol{\beta}^\psi)\right). \tag{14}$$

Remark 4. *We need to point out that the assumptions in Theorem 1 are moderate. For instance, for the assumption that $\nabla\psi(\tilde{\boldsymbol{\beta}}^0) \in \mathrm{col}(X^T)$, we can take $\tilde{\boldsymbol{\beta}}^0 = (\nabla\psi)^{-1}(X^T\mathbf{a})$ for any $\mathbf{a}$. One feasible choice is $\mathbf{a} = \mathbf{0}$, resulting in $\tilde{\boldsymbol{\beta}}_0 = \arg\min_{\boldsymbol{\beta}\in R^p}\psi(\boldsymbol{\beta})$. Since ψ is strongly convex, this minimizer is not hard to calculate, for example: we have*
$\cdot\psi(\cdot) = \|\cdot\|_q^q$ or $\psi(\cdot) = \|\cdot\|_q^2$ for $q > 1 \Rightarrow \arg\min\psi(\cdot) = \mathbf{0}$;
$\cdot\psi(\boldsymbol{\beta}) = \boldsymbol{\beta}^T H\boldsymbol{\beta}$ for a positive definite $H \Rightarrow \arg\min\psi(\cdot) = \mathbf{0}$;
$\cdot\psi(\boldsymbol{\beta}) = \sum_{i=1}^p \beta_i\log(\beta_i) - \beta_i \Rightarrow \arg\min\psi(\cdot) = \mathbf{1}$.

Remark 5. *As for the assumption in (b), we can take $m = \frac{110L\ell n}{\alpha s_m^2}$ and $\eta = \frac{\alpha}{36L}$ to get $\tau' = (1 + 108/110)/2 < 1$. Take a good mirror map such that $\ell/\alpha = \mathcal{O}(1)$ and assume $L/s_m^2 = \mathcal{O}(1)$, we further have $m = \Theta(n)$, so the algorithm can be implemented efficiently.*

It is useful to provide a high level understanding of Theorem 1. It implies that the discrete update of the VRSMD Algorithm on the unregularized objective (2) is an ϵ-solution of the regularized optimization problem (10), so it is an implicit regularization result. Furthermore, since (10) minimizes a strictly convex function over a convex set, the solution will be unique, thus the estimator from VRSMD must converge to this unique solution. The finding in Theorem 1 can be extended to other variants of SMD, which we omit here due to page limit.

IV. Further Understanding of Implicit Regularization

In this section, we provide a deeper understanding of our theoretical results in the previous section by analyzing two special mirror maps for linear regression model: one is $\psi(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2/2$, the other is $\psi(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_{1+\delta}^{1+\delta}$ for small $\delta > 0$.

Let us first consider $\psi(\cdot) = \|\cdot\|_2^2/2$, where VRSMD reduces to SVRG algorithm in [12]. In this case, we further show that $\|\boldsymbol{\beta}^a - \boldsymbol{\beta}^\psi\|_2^2$ linearly converges to 0, where $\boldsymbol{\beta}^\psi$ is the minimum ℓ_2 norm solution $\boldsymbol{\beta}^\psi = (X^TX)^+X^T\mathbf{y} = X^+\mathbf{y}$:

Corollary 1. *Denote $L = \max_i\|\mathbf{x}_i\|_2^2$, take $\psi(\cdot) = \|\cdot\|_2^2/2$, let $\tilde{\boldsymbol{\beta}}^0 = \boldsymbol{\beta}_m^0 = \mathbf{0}$, $\eta_t = \eta < 1/(24L)$, and assume*

$$\tau'' = \frac{12L\eta^2 + n/(ms_m^2)}{\eta - 12L\eta^2} < 1. \tag{15}$$

Run Option II of Algorithm 1 on (2)*, we have*

$$\mathbb{E}\|\boldsymbol{\beta}^a - X^+\mathbf{y}\|_2^2 \le \frac{(\tau'')^S}{s_m^2}\|P_{\mathrm{col}(X)}\mathbf{y}\|_2^2. \tag{16}$$

Next, let us consider the mirror map $\psi(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_{1+\delta}^{1+\delta}$ for a small $\delta > 0$ in VRSMD, which leads to a sparse solution. Assume that $\|\boldsymbol{\beta}_t^s\|_\infty \le K$ for a large enough K throughout the updates, we then have $\psi(\boldsymbol{\beta}_t^s)$ is $\frac{(1+\delta)\delta}{K^{1-\delta}}$-strongly convex and $\|\nabla\psi(\boldsymbol{\beta}^a)\|_2 \le \sqrt{p}(1+\delta)K^\delta$. By Theorem 1 we have the estimator sequence converges to the penalized solution

$$\boldsymbol{\beta}^{(\delta)} = \mathrm{argmin}_{\boldsymbol{\beta}}\left\{\|\boldsymbol{\beta}\|_{1+\delta}^{1+\delta} : X\boldsymbol{\beta} = P_{\mathrm{col}(X)}\mathbf{y}\right\}. \tag{17}$$

This choice of mirror map has $\lim_{\delta\to 0}\|\boldsymbol{\beta}\|_{1+\delta}^{1+\delta} = \|\boldsymbol{\beta}\|_1$ and the solution $\boldsymbol{\beta}^{(0)} := \mathrm{argmin}_{\boldsymbol{\beta}}\{\|\boldsymbol{\beta}\|_1 : X\boldsymbol{\beta} = P_{\mathrm{col}(X)}\mathbf{y}\}$ is sparse. Thus we expect that for small δ, $\boldsymbol{\beta}^{(\delta)}$ is close to $\boldsymbol{\beta}^{(0)}$ and recovers a sparse true parameter.

We now provide a rigorous argument for the sparse recovery. Assume that the data is generated by $\mathbf{y} = X\boldsymbol{\beta}^o$ for a sparse $\boldsymbol{\beta}^o$. In the following theorem we have $\boldsymbol{\beta}^{(\delta)}$ accurately recovers $\boldsymbol{\beta}^o$ when the design matrix X satisfies some proper conditions.

Theorem 2 (Sparse Recovery)**.** *Under the sparse setting defined above, denote $s = \|\boldsymbol{\beta}^o\|_0$. Assume that the design matrix X satisfies (s,γ)-RE condition and is s-good with constant $\kappa < \frac{1}{2}$. For any $\xi > 0$, if we choose*

$$\delta \le \frac{\log\left(1 + \frac{(1-2\kappa)\sqrt{n\gamma}}{\sqrt{s}\|\mathbf{y}\|_2}\xi\right)}{\log p - \log\left(1 + \frac{(1-2\kappa)\sqrt{n\gamma}}{\sqrt{s}\|\mathbf{y}\|_2}\xi\right)}, \tag{18}$$

we have

$$\|\boldsymbol{\beta}^{(\delta)} - \boldsymbol{\beta}^o\|_1 \le \xi. \tag{19}$$

By Theorem 2, the estimator $\boldsymbol{\beta}^{(\delta)}$ estimates the sparse truth with a small error ξ. In this way, VRSMD algorithm 1 achieves near sparse recovery via implicit regularization.

V. Numerical Experiment

Simulation. We generate data by a sparse model as follows: Set $n = 1000, p = 5000 > n$ for the design matrix X. Simulate $X = \Sigma^{1/2}W$ where the entries of W are i.i.d. $N(0,1)$ and $\Sigma = 0.5 * \mathrm{diag}(\mathbf{1}_n) + 0.5 * \mathbf{1}_{n\times n}$. The true parameter $\boldsymbol{\beta}^o \in \mathbb{R}^p$ has its first 30 entries sampled from i.i.d. $N(0,1)$ and the rest entries set to 0. Compute responses $\mathbf{y} = X\boldsymbol{\beta}^o$.

We then run VRSMD on objective function (2) for this simulated X and $\mathbf{y}$. For a range of δ, set the mirror map as $\psi(\cdot) = \|\cdot\|_{1+\delta}^{1+\delta}$, and run VRSMD with initialization $\tilde{\boldsymbol{\beta}}^0 = \mathbf{0}$, step-size $\eta = 0.0002$, outer iteration number 50 and inner iteration number $1000 = n$. The result is in Fig. 1.

Experiment on RNA dataset. We use the gene expression cancer RNA-Seq data set[1] for experiment. The data consists of 801 observations, each of dimension $20,531$. Randomly split the data into 600 training data and 201 testing data. Run VRSMD algorithm on training data using mirror function $\psi = \|\cdot\|_{1.1}^{1.1}$ where initialization $\tilde{\boldsymbol{\beta}}^0 = \mathbf{0}$, step-size $\eta = 0.015$, inner iteration number 400 and outer iteration number determined by 5-fold cross validation (i.e. early stopping). We also compare VRSMD with the Hadamard GD [21, 18], which also has implicit regularization for sparsity. The result is in Fig. 2.

[1] https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq

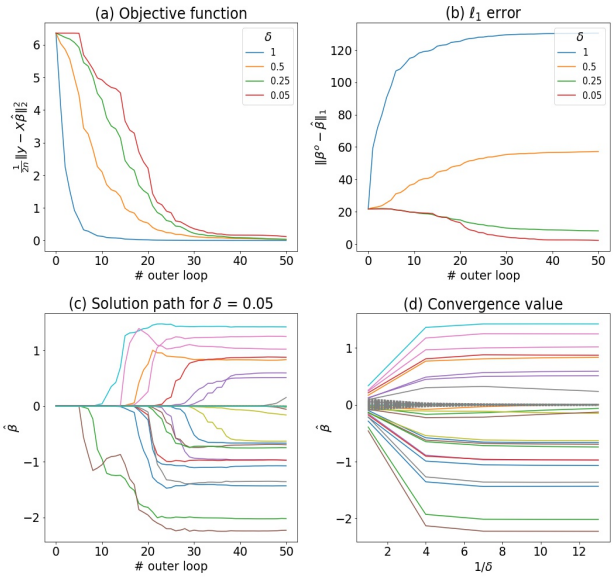

Fig. 1: Run VRSMD Algorithm on simulated noiseless data. (a) the squared error objective function converges quickly to 0. In (b), the ℓ_1 estimation error converges to smaller value for smaller δ, indicating the nearly exact recovery of the sparse signal, which supports Theorem 2. In (c), for $\psi(\cdot) = \|\cdot\|_{1.05}^{1.05}$, the VRSMD estimator converges to the true parameter values. In (d), for a smaller δ, the convergence value of VRSMD estimator is closer to ground truth.

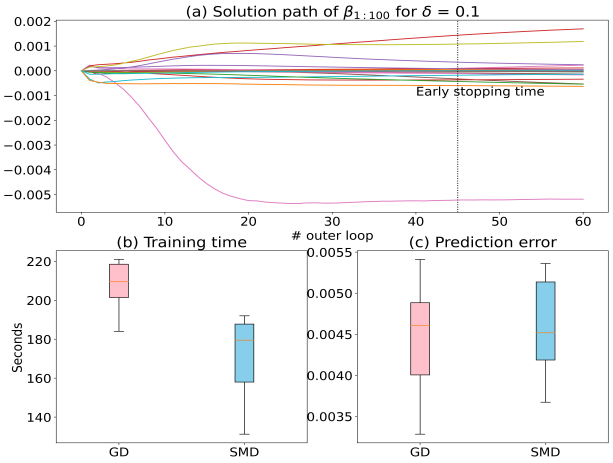

Fig. 2: Run VRSMD Algorithm on RNA dataset. In (a), we plot the solution path of the first 100 entries of $\boldsymbol{\beta}$, and it shows that the early stopped estimator is sparse. In (b) and (c), we compare the performance of VRSMD with Hadamard GD. Plot (b) shows that VRSMD trains faster than Hadamard GD, which is tested significant by one-sided Wilcoxon signed-rank test (p-value = 9.77×10^{-4}). Plot (c) shows that the two algorithms have same prediction error on testing data, which we test by two-sided Wilcoxon signed-rank test (p-value = 0.56).

VI. Conclusions and Future Research

Our work analyzes the implicit regularization property of VRSMD, which covers both underfitting and overfitting cases in linear regression. In particular, our theorem shows that the implicit regularization property can help VRSMD find a sparse ground truth with a small error. Our experiments illustrate that the VRSMD is computationally efficient compared to the Hadamard GD algorithm that has implicit regularization for sparsity [21], thanks to the stochastic nature of VRSMD.

We discuss some future directions of our research. First, it is useful to study the implicit regularization properties of the VRSMD in the nonEuclidean setup. For example, one can consider the generalized linear model (GLM) where the data lies in a Riemannian manifold and mirror descent and/or natural gradient descent are efficient. Our analysis can be extended from the linear regression model to the GLM case.

Second, it is interesting to investigate the minimax property of the VRSMD estimator. We see from our experiment that VRSMD with early stopping has comparable prediction performance to Hadamard GD that is minimax optimal for sparse regression [21, 18]. This leads to the open question of how to select a good mirror map and an optimal stopping time in VRSMD (or variants of VRSMD) such that the resulting estimator is minimax optimal and thus generalizes well.

Third, one can explore the implicit regularization properties of other variants of SMD. For example, one can consider the accelerated version of VRSMD as the Katyusha algorithm in [2]. Such an algorithm is well studies in optimization literature, and it has a better convergence rate that can match the theoretical optimum. We hypothesize that it also enjoys the implicit regularization property, but the proof is out of the scope of this paper.

Finally, our results might provide a better understanding of deep neural networks. By [19], Gradient Descent on Hadamard reparameterized linear regression, which is related to a neural network with multiple layers, can be approximated by Mirror Descent on original parameters. This point of view allows us to study a neural network from the VRSMD perspective and helps to explain why the Gradient Descent gives a sparse estimator in some deep learning models.

Appendix: Proof sketch for Theorem 1

It is easy to check that $\mathbf{v}_t^s \in \mathrm{col}\left(X^T\right)$. Then by $\nabla\psi(\tilde{\boldsymbol{\beta}}^0) \in \mathrm{col}(X^T)$, we immediately have $\nabla\psi\left(\boldsymbol{\beta}^a\right) \in \mathrm{col}\left(X^T\right)$. Combine this key observation with the fact that

$$\psi(\boldsymbol{\beta}^a) - \psi(\boldsymbol{\beta}^\psi) \le \langle \nabla\psi(\boldsymbol{\beta}^a), \boldsymbol{\beta}^a - \boldsymbol{\beta}^\psi \rangle,$$

we have

$$\begin{aligned}
&\mathbb{E}[\psi(\boldsymbol{\beta}^a) - \psi(\boldsymbol{\beta}^\psi)] \le \mathbb{E}\langle \nabla\psi(\boldsymbol{\beta}^a), \boldsymbol{\beta}^a - \boldsymbol{\beta}^\psi \rangle \\
\le & B\mathbb{E}\|P_{\mathrm{col}(X^T)}(\boldsymbol{\beta}^a - \boldsymbol{\beta}^\psi)\|_2 \le \frac{B}{s_m}\sqrt{\mathbb{E}\|X\boldsymbol{\beta}^a - X\boldsymbol{\beta}^\psi\|_2^2} \\
= & \frac{B}{s_m}\sqrt{2n\mathbb{E}(F(\boldsymbol{\beta}^a) - F(\boldsymbol{\beta}^\psi))}. \qquad (20)
\end{aligned}$$

Applying Proposition 1 to (20), we have the desired inequalities.

References

[1] Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. "The implicit regularization of stochastic gradient flow for least squares". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 233–244.

[2] Zeyuan Allen-Zhu. *Katyusha: The First Direct Acceleration of Stochastic Gradient Methods*. 2018. arXiv: 1603.05953 [math.OC].

[3] Zeyuan Allen-Zhu. "Katyusha: The first direct acceleration of stochastic gradient methods". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 8194–8244.

[4] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. "Implicit Regularization in Deep Matrix Factorization". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf.

[5] Navid Azizan and Babak Hassibi. "Stochastic Gradient/Mirror Descent: Minimax Optimality and Implicit Regularization". In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=HJf9ZhC9FX.

[6] Shahar Azulay et al. "On the Implicit Bias of Initialization Shape: Beyond Infinitesimal Mirror Descent". In: *arXiv preprint arXiv:2102.09769* (2021).

[7] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. "Deep learning: a statistical viewpoint". In: *arXiv preprint arXiv:2103.09177* (2021).

[8] Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. "Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process". In: *Proceedings of Thirty Third Conference on Learning Theory*. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 483–513. URL: http://proceedings.mlr.press/v125/blanc20a.html.

[9] Michal Derezinski, Feynman T Liang, and Michael W Mahoney. "Exact expressions for double descent and implicit regularization via surrogate random design". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 5152–5164. URL: https://proceedings.neurips.cc/paper/2020/file/37740d59bb0eb7b4493725b2e0e5289b-Paper.pdf.

[10] Jianqing Fan, Zhuoran Yang, and Mengxin Yu. *Understanding Implicit Regularization in Over-Parameterized Nonlinear Statistical Model*. 2021. arXiv: 2007.08322 [stat.ML].

[11] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. "SPIDER: Near-Optimal Non-Convex Optimization via Stochastic Path-Integrated Differential Estimator". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper/2018/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf.

[12] Rie Johnson and Tong Zhang. "Accelerating stochastic gradient descent using predictive variance reduction". In: *Advances in Neural Information Processing Systems* 26 (2013), pp. 315–323.

[13] Wenjie Li, Zhanyu Wang, Yichen Zhang, and Guang Cheng. "Variance Reduction on Adaptive Stochastic Mirror Descent". In: *arXiv preprint arXiv:2012.13760* (2021).

[14] Yiling Luo, Xiaoming Huo, and Yajun Mei. *Implicit Regularization Properties of Variance Reduced Stochastic Mirror Descent*. 2022. DOI: 10.48550/ARXIV.2205.00058. URL: https://arxiv.org/abs/2205.00058.

[15] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. "SARAH: A Novel Method for Machine Learning Problems Using Stochastic Recursive Gradient". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 2613–2621. URL: http://proceedings.mlr.press/v70/nguyen17b.html.

[16] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. "Stochastic variance reduction for nonconvex optimization". In: *International Conference on Machine Learning*. 2016, pp. 314–323.

[17] Samuel L Smith, Benoit Dherin, David Barrett, and Soham De. "On the Origin of Implicit Regularization in Stochastic Gradient Descent". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=rq_Qr0c1Hyo.

[18] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. "Implicit Regularization for Optimal Sparse Recovery". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/5cf21ce30208cfffaa832c6e44bb567d-Paper.pdf.

[19] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. "The Statistical Complexity of Early-Stopped Mirror Descent". In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 253–264. URL: https://proceedings.neurips.cc/paper/2020/file/024d2d699e6c1a82c9ba986386f4d824-Paper.pdf.

[20] Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. "Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=3X64RLgzY6O.

[21] Peng Zhao, Yun Yang, and Qiao-Chu He. "Implicit regularization via Hadamard product over-parametrization in high-dimensional linear regression". In: *arXiv preprint arXiv:1903.09367* (2019).