Manifold Identification of Dual Averaging Methods for Regularized Stochastic Online Learning

Sangkyun Lee and Stephen J. Wright
Computer Sciences Department,
University of Wisconsin-Madison, USA

THE UNIVERSITY of WISCONSIN MADISON

Introduction

In regularized stochastic learning, we solve

$$\min_{w\in\mathbb{R}^n} \phi(w) := f(w)+\Psi(w), \quad f(w) := \mathbb{E}_\xi F(w;\xi)$$

- $\xi \overset{iid}{\sim} P$, P is supported on $\Xi \subset \mathbb{R}^d$.
- $F(\cdot;\xi) : \mathbb{R}^n \to \mathbb{R}$ is convex, $\forall \xi \in \Xi$.
- $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, and convex.
  - e.g.: $\|w\|_1$ or $\sum_{g\in G} \|w_g\|_2$.
- w^*: a minimizer of $\phi(w)$.
- We use $\|\cdot\|$ to represent $\|\cdot\|_2$.

In online settings,

- At time t, $F(\cdot;\xi_t) + \Psi(\cdot)$ is revealed for $\xi_t \in \Xi$, and
- w_t is made using information gathered so far.
- Goal: to generate $w_1, w_2, \ldots$,:

$$\lim_{t\to\infty} \mathbb{E}\left[F(w_t;\xi) + \Psi(w_t)\right] = f(w^*) + \Psi(w^*).$$

Solution Methods:

SGD (Stochastic Gradient Descent):

$$w_{t+1} = w_t - \eta_t(g_t + h_t), \quad t \geq 1.$$

- $\eta_t = O(1/\sqrt{t})$, or $O(1/t)$ for strongly convex Ψ.
- $g_t \in \partial F(w_t;\xi_t)$ and $h_t \in \partial\Psi(w_t)$.
- Information from Ψ can be combined into $F(\cdot;\xi_t)$.

RDA (Regularized Dual Averaging):

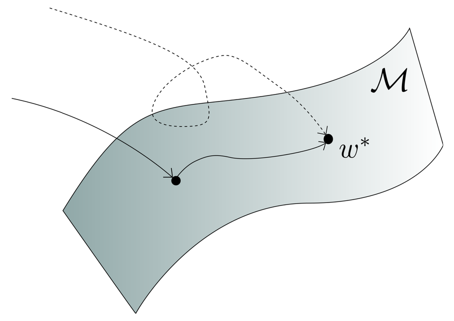
$$w_{t+1} = \arg\min_{w\in\mathbb{R}^n} \left\{ \langle \bar{g}_t, w\rangle + \Psi(w) + \frac{\beta_t}{t}\|w\|^2 \right\}.$$

- By Xiao (2010), extending the primal-dual averaging method (Nesterov, 2009).
- $\beta_t = O(\sqrt{t})$, or $O(1+\ln t)$ for strongly convex Ψ.
- *Dual average*, $\bar{g}_t = \frac{1}{t}\sum_{j=1}^t g_j$, $\quad g_j \in \partial F(w_j;\xi_j)$.
- *Explicit use* of Ψ.

Manifold Identification: Motivation

- RDA finds soln. structures better than SGD,
- but convergence of RDA = SGD.
- Solutions often lie on a low-dim manifold.

The *optimal manifold* is a smooth surface in $\mathbb{R}^n$ containing w^*.



Our contribution:

- ***Proof: RDA identifies the optimal manifold.***
- ***New algorithm RDA$^+$: switches to a rapid convergent optimization method on a near-optimal manifold.***

Assumptions

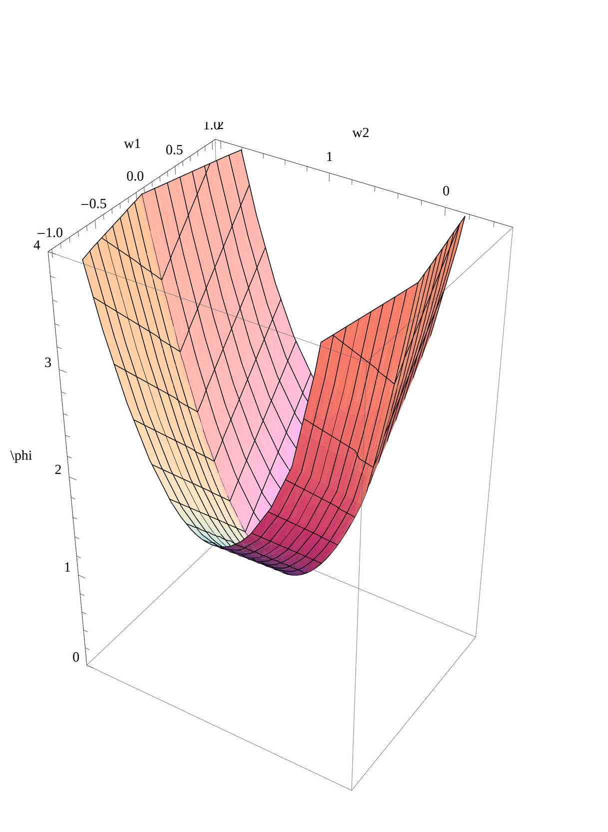
1. *$F(\cdot;\xi)$ is differentiable, for all $\xi \in \Xi$.*
2. *Unbiasedness*:
$$\nabla_w \mathbb{E}_\xi F(w;\xi) = \mathbb{E}_\xi \nabla_w F(w;\xi).$$
3. *Lipschitz Continuity*:
   There exists $L > 0$ such that for all $w, w' \in \mathbb{R}^n$,
$$\|\nabla_w F(w;\xi_t) - \nabla_w F(w';\xi_t)\| \leq L\|w - w'\|.$$
4. *Nondegeneracy of w^*.*
   Optimality: $0 \in \partial\phi(w^*)$,
   Nondegeneracy: $0 \in \mathrm{ri}\,(\partial\phi(w^*))$.
5. *Partial Smoothness of Ψ*:
   Ψ behaves like a smooth function *near w^*, on the optimal manifold $\mathcal{M}$.*
6. *Strong Minimizer Property*:
   w^* is a *strong local minimizer* of ϕ, relative to the optimal manifold $\mathcal{M}$, i.e., there exists $c_\mathcal{M}, r_\mathcal{M} > 0$:
$$\phi|_\mathcal{M}(w) \geq \phi|_\mathcal{M}(w^*) + c_\mathcal{M}\|w - w^*\|^2,$$
$$\forall w \quad \text{s.t. } \|w - w^*\| \leq r_\mathcal{M}.$$

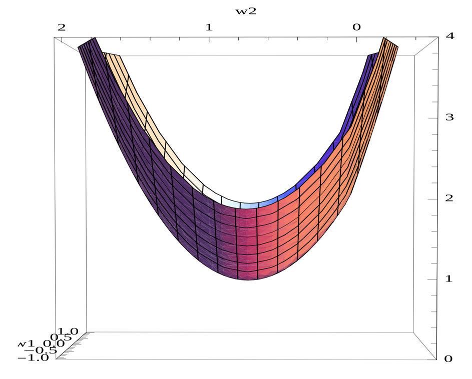
$\Rightarrow$ w^* is a strong local minimizer of ϕ in the full space (Lee and Wright, 2011, Theorem 5): there exist $0 < c < c_\mathcal{M}$ and $0 < \bar{r} < r_\mathcal{M}$:

$$\phi(w) \geq \phi(w^*) + c\|w - w^*\|^2,$$
$$\forall w \quad \text{s.t. } \|w - w^*\| \leq \bar{r}.$$

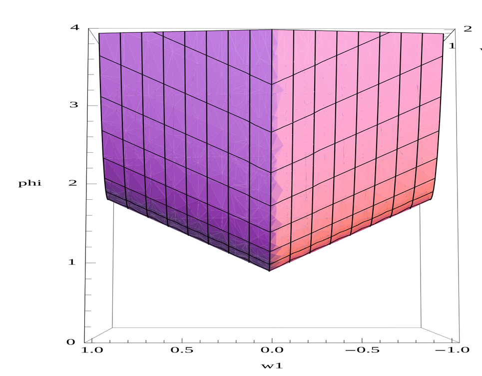
An Example in $\mathbb{R}^2$

$$\phi(w) = f(w) + \Psi(w) = 2(w_2 - 1)^2 + \|w\|_1.$$
$$w^* = (0, 0.75),$$
$$\mathcal{M} = \{w^* + (0,\alpha) \mid \alpha \in (-0.75, \infty)\}.$$





$$\phi|_\mathcal{M}(w) = 2(w_2 - 1)^2 + w_2$$



$$\phi_{\mathcal{M}^\perp}(w) = |w_1| + 0.125$$

Our Results:

Stochastic Behavior of Dual Averages

- $\bar{g}_t$ approaches $\nabla f(w^*)$ in probability.
$$\mathbb{P}(\|\bar{g}_t - \nabla f(w^*)\| > \epsilon) < O(t^{-1/4}),$$

Convergent Sequences

: Majority of w_j from RDA approaches w^* in expectation.

- $I_{(A)} = 1$ if A is true; $I_{(A)} = 0$ otherwise.
- $\mathcal{S}$: the index set of "nice" iterates:

$$\mathcal{S} := \Big\{ j \in \mathbb{N} \mid \mathbb{E}\left[I_{(\|w_j - w^*\| \leq \bar{r})}\|w_j - w^*\|^2\right] \leq j^{-1/4}, \ \&$$
$$\mathbb{E}\left[I_{(\|w_j - w^*\| > \bar{r})}\|w_j - w^*\|\right] \leq \frac{1}{\bar{r}} j^{-1/4} \Big\}.$$

- w_j, $j \in \mathcal{S}$, approaches w^* in probability:
$$\mathbb{P}\left(\|w_j - w^*\| > \epsilon\right) < O(j^{-1/4}), \quad \forall j \in \mathcal{S}.$$
- $\mathcal{S}$ is "dense". For $\mathcal{S}_t := \mathcal{S} \cap \{1,2,\ldots,t\}$,
$$\frac{|\mathcal{S}_t|}{|\{1,2,\ldots,t\}|} > 1 - O(t^{-1/4}),$$

Manifold Identification of RDA

: w_j, $j \in \mathcal{S}$, from RDA eventually lie on the optimal manifold $\mathcal{M}$:

$$\mathbb{P}(w_j \in \mathcal{M}) \geq 1 - O(j^{-1/4}), \quad \text{for suff. large } j \in \mathcal{S}.$$

There is no dependency on the problem dimension n.

RDA$^+$ Algorithm

Initialize: set $w_1 = 0$ and $\bar{g}_0 = 0$.
Dual Averaging:
for $j = 1, 2, \ldots$ **do**
  Choose a random $\xi_j \in \Xi$; $g_j \leftarrow \nabla_w F(w_j;\xi_j)$.
  Update dual average: $\bar{g}_j = \frac{j-1}{j}\bar{g}_{j-1} + \frac{1}{j}g_j$.

$$w_{j+1} = \arg\min_{w\in\mathbb{R}^n} \left\{ \langle \bar{g}_j, w\rangle + \Psi(w) + \frac{\beta_t}{t}\|w\|^2 \right\}.$$

 if $\exists\mathcal{M}$ such that $w_{j+2-i} \in \mathcal{M}$ for $i = 1, 2, \ldots, \tau$ **then**
  Local Phase:
  Expand $\mathcal{M}$ and use LPS (Shi et al., 2008; Wright, 2010) to search for a solution on manifold $\mathcal{M}$, starting at w_{j+1};
 end if
end for

Safeguarding: Expand $\mathcal{M}$ before local phase, by adding components i that may yet contain w^*, i.e.

$$[w_{j+1}]_i = 0 \text{ and } |[\bar{g}_j]_i| > \rho\lambda$$

for some $\rho \in (0, 1]$. Since LPS can find submanifolds of $\mathcal{M}$, this works if $\mathcal{M}$ is a *superset* of the optimal manifold.

Local Phase: Use an approx. obj. with samples in $\mathcal{N}$:

$$\min_{w\in\mathcal{M}} \tilde{\phi}_\mathcal{N}(w) := \tilde{f}_\mathcal{N}(w)+\Psi(w) = \frac{1}{|\mathcal{N}|}\sum_{j\in\mathcal{N}} F(w;\xi_j)+\Psi(w)$$

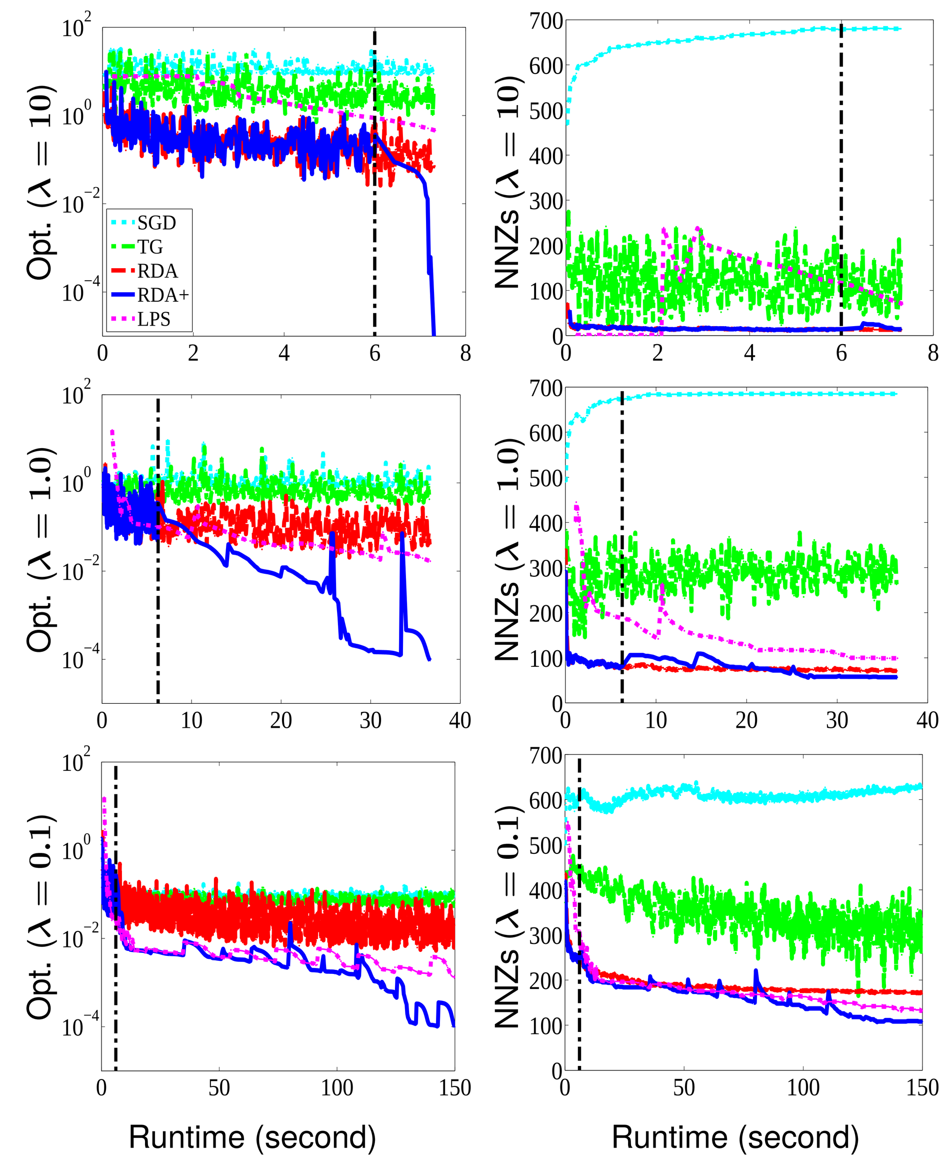
Optimality measure: $\delta(w_j) := \frac{1}{\sqrt{n}} \inf_{a_j \in \partial\Psi(w_j)} \|\nabla \tilde{f}_\mathcal{N}(w_j) + a_j\|.$

Experiments

- Binary classification via ℓ_1-regularized logistic regression.
$$F(w;\xi_t) = \log\left(1 + \exp\left(-y_t w^T x_t\right)\right),$$
$$\Psi(w) = \lambda\|w\|_1, \ \lambda > 0.$$
- MNIST 6 vs. 7: 12183 training / 1986 test.
- $\tau = 100$ (max it. = 19327), $\delta < 10^{-4}$, $\rho = 0.85$, and $\mathcal{N}$ = full training set.
- Compared RDA$^+$ to SGD, TG (Langford et al., 2009), RDA, and LPS (batch method).

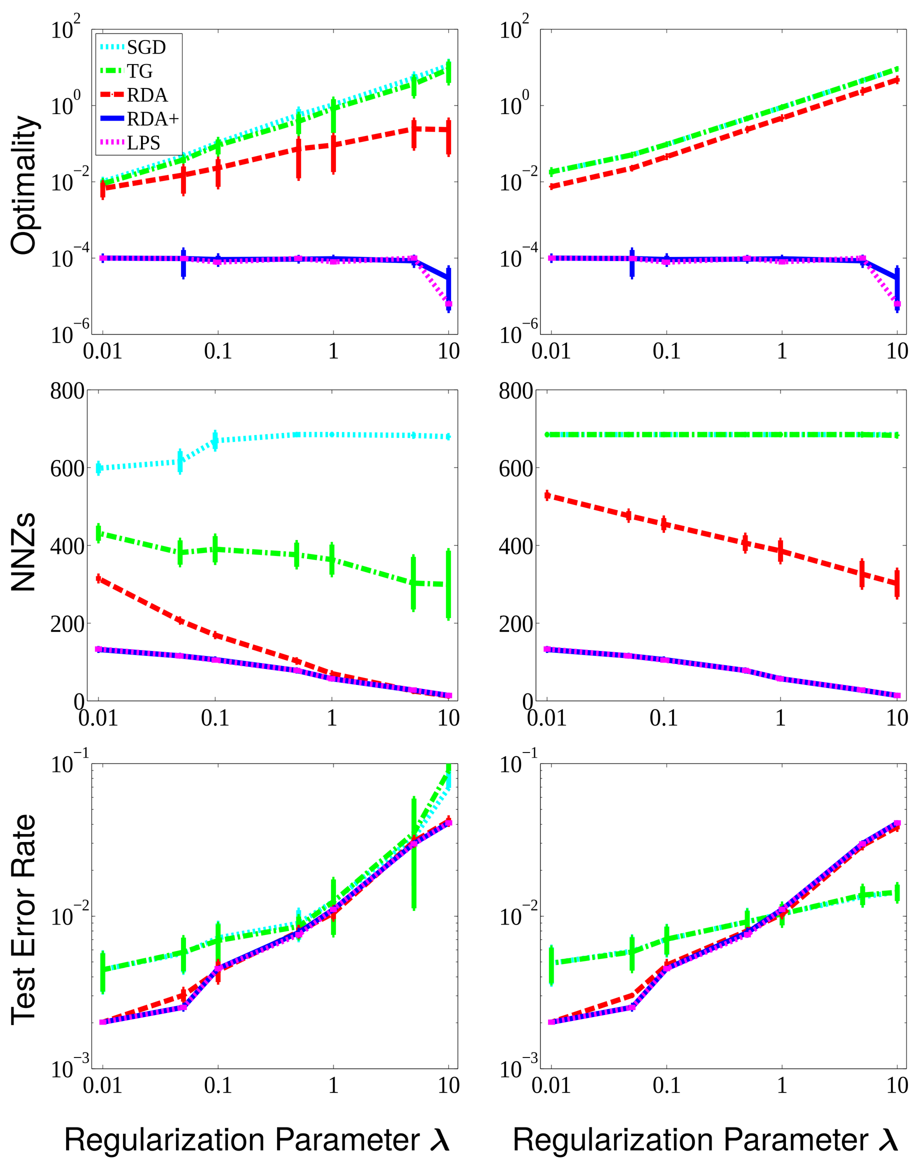
Progress in Time

- Optimality:
  - RDA$^+$ achieves target opt. faster than others.
  - RDA behaves better than SGD and TG, but it hardly achieves the target value.
- NNZs:
  - RDA: sparser solns. with less fluctuation than SGD and TG.
  - RDA: fails to identify the smallest NNZ set of RDA$^+$ in given time.
- LPS vs RDA$^+$ (local):
  - local phase often converges faster than LPS; operating on reduced spaces.



Quality of Solutions

- Settings:
  - 100 repetitions.
  - LPS: no time limit (about $\times 4$ runtime of RDA$^+$.)
  - No primal averaging in RDA$^+$ and LPS (duplicated results on the right).
- Optimality:
  - Only RDA$^+$ achieves the target optimality and smallest NNZs.
  - RDA$^+$: almost identical quality to LPS.
- NNZ:
  - RDA: similar NNZs to RDA$^+$ for large λ, but not on smaller values.
- Test error rate:
  - RDA$^+$ shows small improvement: this is marginal, but hard to achieve solely with SGD-type methods in limited time.



sklee@cs.wisc.edu, http://pages.cs.wisc.edu/~sklee

Sangkyun Lee and Stephen J. Wright
Computer Sciences Department,
University of Wisconsin-Madison, USA

J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, 10:777–801, 2009.
S. Lee and S. J. Wright. Manifold identification of dual averaging algorithm for regularized stochastic online learning. Technical report, University of Wisconsin-Madison, April 2011.
Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009.
W. Shi, G. Wahba, S. J. Wright, K. Lee, R. Klein, and B. Klein. LASSO-Patternsearch algorithm with application to opthalmology data. *Statistics and its Interface*, 1:137–153, January 2008.
S. J. Wright. Accelerated block-coordinate relaxation for regularized optimization. Technical report, University of Wisconsin-Madison, August 29 2010.
L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, October 2010.

sklee@cs.wisc.edu, http://pages.cs.wisc.edu/~sklee