A Varying-Coefficient Regularized Dual Averaging Alg. for Regularized Stochastic Optimization

Shiqian Ma

Department of Mathematics, UC Davis

Joint Work with Conghui Tan (CUHK) and Tong Zhang (Tencent AI Lab)

Bay Area Optimization Meeting
Stanford University
May 19, 2018
Regularized stochastic optimization

\[
\min_{x \in X} \left\{ F(x) \triangleq \mathbb{E}_\xi [f(x, \xi)] + \psi(x) \right\},
\]

- \(X\): bounded convex set
- \(f(x, \xi)\): convex loss function wrt sample data \(\xi\), we don’t assume smoothness of \(f\)
- \(\psi\): convex regularizer
Empirical Risk Minimization

- Supervised learning
  \[
  \min_{x \in \mathcal{X}} \left\{ F(x) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(x, \xi_i) + \psi(x) \right\},
  \]
  where \( \xi_i \) denotes the \( i \)-th data sample.

- Example: sparse logistic regression
  \[
  \min_{x} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log (1 + \exp(-b_i \langle x, a_i \rangle)) + \lambda \| x \|_1 \right\},
  \]
  where \( a_i \in \mathbb{R}^d \) and \( b_i \in \{ \pm 1 \} \) are the feature vector and label of the \( i \)-th sample.

- Example: hinge loss sparse SVM
  \[
  \min_{x} \left\{ \frac{1}{n} \sum_{i=1}^{n} \max\{1 - b_i \langle a_i, x \rangle, 0\} + \lambda \| x \|_1 \right\}.
  \]
Stochastic Gradient Descent

\[
\min_{x \in X} \left\{ F(x) \triangleq \mathbb{E}_\xi [f(x, \xi)] + \psi(x) \right\},
\]

- If \( \psi \) vanishes, then SGD
  \[
  x^{t+1} \leftarrow \arg\min_{x \in X} \left\{ \langle x, g^t \rangle + \frac{1}{2\alpha_t} \|x - x^t\|^2_2 \right\},
  \]
  
- \( g^t \in \partial f(x^t, \xi^t) \) and \( \xi^t \) is randomly sampled.

- \( \alpha_t \): step size.

- Non-ergodic convergence rate: \( O(\ln t/\sqrt{t}) \) for convex \( F \) and \( O(\ln t/t) \) for strongly convex \( F \) \( (\text{Shamir-Zhang-2013}) \)

- Ergodic convergence rate: \( O(1/\sqrt{t}) \) for convex \( F \) and \( O(1/t) \) for strongly convex \( F \) \( (\text{Shamir-Zhang-2013}) \)
When $\psi$ presents, proximal SGD

$$x^{t+1} \leftarrow \arg \min_{x \in \mathcal{X}} \left\{ \langle x, g^t \rangle + \psi(x) + \frac{1}{2\alpha_t} \|x - x^t\|^2 \right\},$$

$g^t \in \partial f(x^t, \xi^t)$ and $\xi^t$ is randomly sampled.

Ergodic rate of PSGD: $O(1/\sqrt{t})$ for convex $F$, and $O(\ln t/t)$ for strongly convex $F$ (Duchi, et al. 2010)
Regularized Dual Averaging

- Regularized Dual Averaging (RDA) (Xiao-2010)

\[
x^{t+1} \leftarrow \arg \min_{x \in \mathcal{X}} \left\{ \left\langle x, \sum_{k=0}^{t} g^k \right\rangle + (t + 1)\psi(x) + \frac{1}{2\beta_t} \|x - x^0\|_2^2 \right\},
\]

- \( g^t \in \partial f(x^t, \xi^t) \).

- Ergodic rate of RDA: \( \mathcal{O}(1/\sqrt{t}) \) for convex \( F \), and \( \mathcal{O}(\ln t/t) \) for strongly convex \( F \). (Xiao-2010)
Compare PSGD and RDA

- PSGD can be rewritten as

$$x^{t+1} \leftarrow \arg \min_{x \in X} \left\{ \alpha_t \psi(x) + \frac{1}{2} \left\| x - (x^t - \alpha_t g^t) \right\|^2 \right\},$$

- RDA can be rewritten as

$$x^{t+1} \leftarrow \arg \min_{x \in X} \left\{ (t + 1) \beta_t \psi(x) + \frac{1}{2} \left\| x - \left( x^0 - \beta_t \sum_{k=0}^{t} g^k \right) \right\|^2 \right\}.$$

- In a special case, if $\psi(x) \equiv 0$, $X = \mathbb{R}^d$, and $\alpha_t = \beta_t$ are both constants, then RDA and SGD are equivalent.
Further compare PSGD and RDA (McMahan-2011)

- PSGD can be rewritten as

\[
x^{t+1} \leftarrow \arg\min_{x \in X} \left\{ \langle x, \sum_{k=0}^{t} g_k \rangle + \langle x, \sum_{k=0}^{t} \phi_k \rangle + \psi(x) \\ + \frac{1}{2} \sum_{k=0}^{t} \| x - x^k \|^2 \right\},
\]

where \( \phi_k \in \partial \psi(x^{k+1}) \)

- RDA can be rewritten as

\[
x^{t+1} \leftarrow \arg\min_{x \in X} \left\{ \langle x, \sum_{k=0}^{t} g_k \rangle + (t + 1) \psi(x) + \frac{1}{2} \sum_{k=0}^{t} \| x - x^0 \|^2 \right\}.
\]
RDA has two main advantages over PSGD:

- Stronger regularization effect. If $\psi(x) = \lambda \|x\|_1$, then solution more sparse
- RDA can be more effectively implemented on sparse data: it can update the iterate incrementally with only $O(\|g^t\|_0)$ operations at iteration $t$, while PSGD requires $O(d)$ operations

However, some experimental evidence shows that PSGD is faster than RDA in many settings.

**Question:** How to combine the advantages of RDA and PSGD?
Proximal-FTRL

- Proximal follow-the-regularized-leader (FTRL-Proximal) algorithm (McMahan-2011)

\[
x^{t+1} \leftarrow \arg\min_{x \in X} \left\{ \left\langle x, \sum_{k=0}^{t} g^k \right\rangle + (t + 1)\psi(x) + \sum_{k=0}^{t} \frac{\gamma_k}{2} \|x - x^k\|_2^2 \right\}.
\]

- FTRL-Proximal inherits the advantages of both RDA and PSGD:
  - Fast convergence speed (empirically)
  - Ability to induce sparsity (empirically)

- When \( \psi(x) \equiv 0, X = \mathbb{R}^d \), FTRL-Proximal, PSGD, and RDA are all equivalent.

- Ergodic rate of FTRL-Proximal: \( O(1/\sqrt{t}) \) for convex problems. (McMahan-2017)
Our contributions

- A new variant of RDA: Varying-Coefficient RDA
- Shares the advantages of both RDA and PSGD.
- Similar convergence rate as RDA and PSGD.
Varying-Coefficient RDA

- **VC-RDA** (motivation: assign different weights to $g_k$)

  \[
  x^{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} \left\{ \left\langle x, \sum_{k=0}^{t} \alpha_k g^k \right\rangle + \left( \sum_{k=0}^{t} \alpha_k \right) \cdot \psi(x) + \frac{1}{2} \|x - x^0\|^2 \right\},
  \]

- **RDA**

  \[
  x^{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} \left\{ \left\langle x, \sum_{k=0}^{t} g^k \right\rangle + (t + 1)\psi(x) + \frac{1}{2\beta_t} \|x - x^0\|^2 \right\},
  \]

- **PSGD**

  \[
  x^{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} \left\{ \langle x, g^t \rangle + \psi(x) + \frac{1}{2\alpha_t} \|x - x^t\|^2 \right\},
  \]

- **FTRL-Proximal**

  \[
  x^{t+1} \leftarrow \arg\min_{x \in \mathcal{X}} \left\{ \left\langle x, \sum_{k=0}^{t} g^k \right\rangle + (t + 1)\psi(x) + \sum_{k=0}^{t} \frac{\gamma_k}{2} \|x - x^k\|^2 \right\}.
  \]
Algorithm VC-RDA

- Initialize $x^0 = \arg\min_{x \in X} \psi(x)$
- Set $s^{-1} = 0$ and $A_{-1} = 0$
- For $t = 0, 1, \ldots$
  - randomly sample $\xi^t$ and compute a subgradient $g^t \in \partial f(x^t, \xi^t)$
  - Update the sum of subgradients and weights:
    \[
    s^t = s^{t-1} + \alpha_t g^t \\
    A_t = A_{t-1} + \alpha_t
    \]
- Compute the next iterate:
  \[
  x^{t+1} \leftarrow \arg\min_{x \in X} \left\{ \langle x, s^t \rangle + A_t \cdot \psi(x) + B_h(x, x^0) \right\}
  \]
VC-RDA with Adaptive Diagonal Scaling (Ada-VC-RDA)

(Similar to AdaGrad (Duchi-et al.-2011))

- Initialize $x_0 = \text{argmin}_{x \in X} \psi(x)$
- Set $q^{-1} = 0$, $s^{-1} = 0$ and $A_{-1} = 0$
- For $t = 0, 1, \ldots$
  - Randomly sample $\xi^t$ and compute a subgradient $g^t \in \partial f(x^t, \xi^t)$
  - Update the sum of subgradients and weights:
    \[
    s^t = s^{t-1} + \alpha_t g^t \\
    q_i^t = \sqrt{(q_i^{t-1})^2 + (\alpha_t g_i^t)^2 / (A_{t-1}\mu + 1)}, \quad \forall i \\
    A_t = A_{t-1} + \alpha_t
    \]
- Let $Q^t = \text{diag} \left( q^t / (\max_i q_i^t) \right)$
- Compute the next iterate:
  \[
  x^{t+1} \leftarrow \text{argmin}_{x \in X} \left\{ \langle x, s^t \rangle + A_t \cdot \psi(x) + B_h(Q^t x, Q^t x^0) \right\}
  \]
Convergence results

Theorem: Convergence of VC-RDA

Assume \( \{\alpha_t\} \) decreasing. Define \( \bar{x}^t = \sum_{k=0}^{t} \frac{\alpha_k}{A_t} x^k \). VC-RDA satisfies

\[
\mathbb{E} \left[ F(\bar{x}^t) - F(x^*) \right] \leq \frac{D^2}{A_t} + \frac{1}{A_t} \sum_{k=0}^{t} \frac{G^2 \alpha_k^2}{2(A_k - 1 \mu + 1)}.
\]

Corollary

- If \( \psi(x) \) is strongly convex, and \( \alpha_t = \alpha \), VC-RDA satisfies

\[
\mathbb{E} \left[ F(\bar{x}^t) - F(x^*) \right] \leq \frac{2D^2 + \alpha G^2 + \alpha \mu^{-1} G^2 \ln(\mu \alpha t + 1)}{2 \alpha (t + 1)} = O \left( \frac{\ln t}{t} \right).
\]

- If \( \psi(x) \) is non-strongly convex, set \( \alpha_t = \alpha / \sqrt{t + 1} \), then

\[
\mathbb{E} \left[ F(\bar{x}^t) - F(x^*) \right] \leq \frac{2D^2 + \alpha^2 G^2 \ln(t + 1)}{2 \alpha \sqrt{t + 1}} = O \left( \frac{\ln t}{\sqrt{t}} \right).
\]
Theorem: Convergence of Ada-VC-RDA

Assume $\psi(x)$ and $h(x)$ have separable structures, i.e.,
$\psi(x) = \sum_{i=1}^{d} \psi_i(x_i)$ and $h(x) = \sum_{i=1}^{d} h_i(x_i)$. Ada-VC-RDA satisfies

$$
\mathbb{E} \left[ F(\tilde{x}^t) - F(x^*) \right] \leq \frac{D_t^2}{A_t} + \frac{\max_j q_j^t}{A_t} \cdot \sum_{i=1}^{d} q_i^t,
$$

where $D_t^2 = \sup_{x \in X} B_h(Q^t x, Q^t x^0)$. This is approximately

$$
O \left( \frac{\ln t}{\sqrt{t}} \right)
$$
Experiments

- $\ell_1$-regularized logistic regression

$$
\min_x \left\{ \frac{1}{n} \sum_{i=1}^{n} \log (1 + \exp(-b_i \langle x, a_i \rangle)) + \lambda \|x\|_1 \right\},
$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \{\pm 1\}$ are the feature vector and label of the $i$-th sample.

- Note that the loss function is differentiable.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># examples</th>
<th># features</th>
<th>Prop. nonzero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-Sim</td>
<td>72,309</td>
<td>20,958</td>
<td>0.245%</td>
</tr>
<tr>
<td>RCV1</td>
<td>804,414</td>
<td>47,236</td>
<td>0.157%</td>
</tr>
<tr>
<td>URL</td>
<td>2,396,130</td>
<td>3,231,961</td>
<td>0.004%</td>
</tr>
</tbody>
</table>

Table: Statistics of Datasets
Figure: Results on three different datasets. $\lambda = 10^{-4}$ in all problems.
Effects of regularizer

**Figure:** Convergence rate under different choice of $\lambda$ on RCV1 dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RCV1</th>
<th>Real-Sim</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>$10^{-3}$</td>
<td>$10^{-4}$</td>
<td>$10^{-5}$</td>
</tr>
<tr>
<td>Ada-VC-RDA</td>
<td>$8.3e-4$</td>
<td>$7.0e-3$</td>
<td>$3.3e-2$</td>
</tr>
<tr>
<td>FTRL-Prox</td>
<td>$8.3e-4$</td>
<td>$7.2e-3$</td>
<td>$3.2e-2$</td>
</tr>
<tr>
<td>RDA</td>
<td>$8.9e-4$</td>
<td>$9.9e-3$</td>
<td>$7.2e-2$</td>
</tr>
<tr>
<td>PSGD</td>
<td>$3.7e-2$</td>
<td>$9.8e-2$</td>
<td>$2.1e-1$</td>
</tr>
</tbody>
</table>
Figure: Comparison on non-adaptive algorithms. Ada-VC-RDA is still reported as a baseline. $\lambda = 10^{-4}$ for all datasets. The same step sizes are used for all non-adaptive methods.
Nonsmooth Loss Functions

Sparse support vector machine (SSVM):

\[
\min_x \left\{ \frac{1}{n} \sum_{i=1}^{n} \max\{1 - b_i \langle a_i, x \rangle, 0\} + \lambda \|x\|_1 \right\}.
\]

Solution sub-optimality

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RCV1</th>
<th>Real-Sim</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda)</td>
<td>(10^{-3})</td>
<td>(10^{-4})</td>
<td>(10^{-5})</td>
</tr>
<tr>
<td>Ada-VC-RDA</td>
<td>3.5e-4</td>
<td>1.4e-3</td>
<td>2.4e-3</td>
</tr>
<tr>
<td>FTRL-Prox</td>
<td>7.2e-4</td>
<td>1.6e-3</td>
<td>6.6e-3</td>
</tr>
<tr>
<td>RDA</td>
<td>4.4e-4</td>
<td>3.1e-3</td>
<td>8.9e-3</td>
</tr>
<tr>
<td>PSGD</td>
<td>2.7e-2</td>
<td>9.9e-3</td>
<td>6.0e-3</td>
</tr>
</tbody>
</table>

Solution sparsity

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RCV1</th>
<th>Real-Sim</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda)</td>
<td>(10^{-3})</td>
<td>(10^{-4})</td>
<td>(10^{-5})</td>
</tr>
<tr>
<td>Ada-VC-RDA</td>
<td>2.4e-3</td>
<td>1.3e-2</td>
<td>5.8e-2</td>
</tr>
<tr>
<td>FTRL-Prox</td>
<td>2.5e-3</td>
<td>1.3e-2</td>
<td>5.1e-2</td>
</tr>
<tr>
<td>RDA</td>
<td>2.5e-3</td>
<td>1.6e-2</td>
<td>1.1e-1</td>
</tr>
<tr>
<td>PSGD</td>
<td>1.1e-1</td>
<td>1.2e-1</td>
<td>5.7e-1</td>
</tr>
</tbody>
</table>
AUC: RCV1

![AUC vs Sparsity Graph]

- VC_RDA
- FTRL
- RDA

Sparsity

AUC
AUC: Real-Sim

![Graph showing the AUC vs. Sparsity for VC_RDA, FTRL, and RDA. The graph plots the AUC on the y-axis against sparsity on the x-axis. The data suggests that the AUC increases with increasing sparsity for all methods, with VC_RDA consistently higher than FTRL and RDA.]
Thank you for your attention!