Composite Objective Mirror Descent

John C. Duchi\textsuperscript{1,3} Shai Shalev-Shwartz\textsuperscript{2} Yoram Singer\textsuperscript{3} Ambuj Tewari\textsuperscript{4}

\textsuperscript{1}University of California, Berkeley
\textsuperscript{2}Hebrew University of Jerusalem, Israel
\textsuperscript{3}Google Research
\textsuperscript{4}Toyota Technological Institute, Chicago

June 29, 2010
Large scale logistic regression

Problem: \( n \) huge,

\[
\min_x \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(\langle a_i, x \rangle)) + \lambda \|x\|_1
\]

\[= f(x)\]

“Usual” approach: online gradient descent (Zinkevich ’03). Let

\[g_t = \nabla \log(1 + \exp(\langle a_t, x_t \rangle))\]

\[x_{t+1} = x_t - \eta_t g_t - \eta_t \lambda \text{sign}(x_t)\]

Then perform online to batch conversion
Problems with usual approach

- Regret bound/convergence rate: set
  \[ G = \max_t \| g_t + \lambda \text{sign}(x_t) \|_2 \]

  \[
  f(x_T) + \lambda \| x_T \|_1 = f(x^*) + \lambda \| x^* \|_1 + O \left( \frac{\| x^* \|_2 G}{\sqrt{T}} \right)
  \]

  But \( G = \Theta(\sqrt{d}) \)—additional penalty of \( \text{sign}(x_t) \)

- No sparsity in \( x_T \)
Problems with usual approach

- Regret bound/convergence rate: set
  \[
  G = \max_t \| g_t + \lambda \text{sign}(x_t) \|_2
  \]

  \[
  f(x_T) + \lambda \| x_T \|_1 = f(x^*) + \lambda \| x^* \|_1 + O \left( \frac{\| x^* \|_2 G}{\sqrt{T}} \right)
  \]

  But \( G = \Theta(\sqrt{d}) \)—additional penalty of \( \text{sign}(x_t) \)

- No sparsity in \( x_T \)

- Why should we suffer from \( \| \cdot \|_1 \) term?
Online Gradient Descent

Let $g_t = \nabla \log(1 + \exp(\langle a_t, x_t \rangle)) + \lambda \text{sign}(x_t)$. OGD step (Zinkevich '03):

$$x_{t+1} = x_t - \eta g_t = \arg\min_x \left\{ \eta \langle g_t, x \rangle + \frac{1}{2} \| x - x_t \|^2 \right\}$$

$$f(x) + \lambda \|x\|_1$$

$$f(x_t) + \langle g_t, x - x_t \rangle$$
Online Gradient Descent

Let $g_t = \nabla \log(1 + \exp(\langle a_t, x_t \rangle)) + \lambda \text{sign}(x_t)$. OGD step (Zinkevich '03):

$$x_{t+1} = x_t - \eta g_t = \arg\min_x \left\{ \eta \langle g_t, x \rangle + \frac{1}{2} \| x - x_t \|^2 \right\}$$

$$\langle g_t, x \rangle + B_\psi(x, x_t)$$
Problems with Subgradient Methods

- Subgradients are non-informative at singularities
Problems with Subgradient Methods

- Subgradients are non-informative at singularities
Problems with Subgradient Methods

- Subgradients are non-informative at singularities
Composite Objective Approach

Let $g_t = \nabla \log(1 + \exp(\langle a_t, x_t \rangle))$. Truncated gradient (Langford et al. ’08, Duchi & Singer ’09):

$$x_{t+1} = \arg\min_x \left\{ \frac{1}{2} \| x - x_t \|^2 + \eta \langle g_t, x \rangle + \eta \lambda \| x \|_1 \right\}$$

$$= \text{sign}(x_t - \eta g_t) \odot [\| x_t - \eta g_t \| - \eta \lambda]_+$$
Composite Objective Approach

Update is

\[ x_{t+1} = \text{sign}(x_t - \eta g_t) \odot [ |x_t - \eta g_t| - \eta \lambda ]_+ \]

Two nice things:

- Sparsity from \([\cdot]_+\)
- Convergence rate: let \( G = \max_t \|g_t\|_2 \)

\[
f(x_T) + \lambda \|x_T\|_1 = f(x^*) + \lambda \|x^*\|_1 + O\left( \frac{\|x^*\|_2 G}{\sqrt{T}} \right)
\]

No extra penalty from \( \lambda \|x\|_1 \)!
Repeat:

- Learner plays point $x_t$
- Receive $f_t + \varphi$ ($\varphi$ known)
- Suffer loss $f_t(x_t) + \varphi(x_t)$

Goal: attain small regret

$$R(T) := \sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - \inf_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x) + \varphi(x)$$
Composite Objective Mirror Descent

Let $g_t = \nabla f_t(x_t)$. COMID step:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ B_\psi(x, x_t) + \eta \langle g_t, x \rangle + \eta \varphi(x) \right\}$$
Composite Objective Mirror Descent

Let $g_t = \nabla f_t(x_t)$. COMID step:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \{B_\psi(x, x_t) + \eta \langle g_t, x \rangle + \eta \varphi(x)\}$$
Convergence Results

Old (online gradient/mirror descent):

**Theorem:** For any $x^* \in \mathcal{X}$,

$$
\sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - f_t(x^*) - \varphi(x^*) 
\leq \frac{1}{\eta} B_{\psi}(x^*, x_1) + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f_t(x_t) + \nabla \varphi(x_t)\|_*^2
$$
Convergence Results

Old (online gradient/mirror descent):

**Theorem:** For any $x^* \in \mathcal{X}$,

$$
\sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - f_t(x^*) - \varphi(x^*) 
\leq \frac{1}{\eta} B_\psi(x^*, x_1) + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f_t(x_t) + \nabla \varphi(x_t)\|_*^2
$$

New (**COMID**):

**Theorem:** For any $x^* \in \mathcal{X}$,

$$
\sum_{t=1}^{T} f_t(x_t) + \varphi(x_t) - f_t(x^*) - \varphi(x^*) 
\leq \frac{1}{\eta} B_\psi(x^*, x_1) + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f_t(x_t)\|_*^2
$$
Derived Algorithms

- FOBOS (Duchi & Singer, 2009)
- \( p \)-norm divergences
- Mixed-norm regularization
- Matrix COMID
$p$-norms

Better $\ell_1$ algorithms:

$$\varphi(x) = \lambda \|x\|_1$$
Better $\ell_1$ algorithms:

$$\varphi(x) = \lambda \|x\|_1$$

- Idea: non-Euclidean geometry (e.g. dense gradients, sparse $x^*$)
- Recall $\frac{1}{2(p-1)} \|x\|_p^2$ is strongly convex over $\mathbb{R}^d$ w.r.t. $\ell_p$, $1 < p \leq 2$
- Take $\psi(x) = \frac{1}{2} \|x\|_p^2$

**Corollary:** When $\|f'_t(x_t)\|_\infty \leq G_\infty$, take $p = 1 + 1/\log d$ to get

$$R(T) = O \left( \|x^*\|_1 G_\infty \sqrt{T \log d} \right)$$
Derived $p$-norm algorithms

SMIDAS (Shalev-Shwartz & Tewari 2009): take $\varphi(x) = \lambda \|x\|_1$. Assume $\text{sign}([\nabla \psi(x)]_j) = \text{sign}(x_j)$, define

$$S_\lambda(z) = \text{sign}(z) \cdot [|z| - \lambda]_+$$

Then

$$x_{t+1} = (\nabla \psi)^{-1} S_{\eta \lambda} (\nabla \psi(x_t) - \eta f'_t(x_t))$$
**COMID with mixed norms**

\[
\varphi(X) = \|X\|_{\ell_1/\ell_q} = \sum_{j=1}^{d} \|\overline{x}_j\|_q
\]

\[
X = \begin{bmatrix}
\overline{x}_1 \\
\overline{x}_2 \\
\vdots \\
\overline{x}_d
\end{bmatrix} \Rightarrow \begin{bmatrix}
\|\overline{x}_1\|_q \\
\|\overline{x}_2\|_q \\
\vdots \\
\|\overline{x}_d\|_q
\end{bmatrix}
\]

- Separable and solvable using previous methods
- Multitask and multiclass learning
  - \(\overline{x}_j\) associated with feature \(j\)
  - Penalize \(\overline{x}_j\) once
Mixed-norm $p$-norm algorithms

Specialize problem to

$$\min_x \langle v, x \rangle + \frac{1}{2} \|x\|^2_p + \lambda \|x\|_\infty$$

- Closed form? No.
Mixed-norm $p$-norm algorithms

Specialize problem to

$$\min_x \langle v, x \rangle + \frac{1}{2} \|x\|_p^2 + \lambda \|x\|_\infty$$

- Closed form? No.
- Dual problem ($x^* = v - \beta$):

$$\min_{\beta} \|v - \beta\|_q \quad \text{subject to} \quad \|\beta\|_1 \leq \lambda$$
Mixed-norm $p$-norm algorithms

Problem:

$$\min_{\beta} \|v - \beta\|_q \quad \text{subject to} \quad \|\beta\|_1 \leq \lambda$$

Observation: Monotonicity of $\beta$, so $v_i \geq v_j$ implies $\beta_i \geq \beta_j$
Mixed-norm $p$-norm algorithms

Problem:

$$\min_{\beta} \| \nu - \beta \|_q \quad \text{subject to} \quad \| \beta \|_1 \leq \lambda$$

Observation: Monotonicity of $\beta$, so $\nu_i \geq \nu_j$ implies $\beta_i \geq \beta_j$

Root-finding problem:

$$\lambda = \sum_{i=1}^{d} \beta_i(\theta) = \sum_{i=1}^{d} \left[ \nu_i - \theta^{1/(q-1)} \right]_+$$

Solve with median-like search
Matrix COMID

Idea: get sparsity in spectrum of $X \in \mathbb{R}^{d_1 \times d_2}$. Take

$$\varphi(X) = \|X\|_1 = \sum_{i=1}^{\min\{d_1, d_2\}} \sigma_i(X)$$
Matrix COMID

Idea: get sparsity in spectrum of $X \in \mathbb{R}^{d_1 \times d_2}$. Take

$$\varphi(X) = \|X\|_1 = \sum_{i=1}^{\min\{d_1,d_2\}} \sigma_i(X)$$

Schatten $p$-norms: apply $p$-norms to columns of $X \in \mathbb{R}^{d_1 \times d_2}$

$$\|X\|_p = \|\sigma(X)\|_p = \left( \sum_{i=1}^{\min\{d_1,d_2\}} \sigma_i(X)^p \right)^{1/p}$$

Important fact: for $1 < p \leq 2$,

$$\psi(X) = \frac{1}{2(p - 1)} \|X\|_p^2$$

is strongly convex w.r.t. $\|\cdot\|_p$ (Ball et al., 1994)
Matrix **COMID**

Schatten \( p \)-norms: apply \( p \)-norms to columns of \( X \in \mathbb{R}^{d_1 \times d_2} \)

\[
\|X\|_p = \|\sigma(X)\|_p = \left( \sum_{i=1}^{\min\{d_1,d_2\}} \sigma_i(X)^p \right)^{1/p}
\]

Important fact: for \( 1 < p \leq 2 \),

\[
\psi(X) = \frac{1}{2(p-1)} \|X\|_p^2
\]

is strongly convex w.r.t. \( \| \cdot \|_p \) (Ball et al., 1994)

**Consequence:** Take \( p = 1 + 1/\log d \), \( G_\infty \geq \|f'_t(X_t)\|_\infty \). **COMID** with above \( \psi \) has

\[
R(T) = O \left( G_\infty \|X^*\|_1 \sqrt{T \log d} \right)
\]
Trace-norm Regularization

Idea: get sparsity in spectrum, take $\varphi(X) = \|X\|_1 = \sum_i \sigma_i(X)$

$$X_{t+1} = \arg\min_{X \in \mathcal{X}} \eta \langle f'_t(X_t), X \rangle + B_\psi(X, X_t) + \eta \lambda \|X\|_1$$

For $1 < p \leq 2$, update is

Compute SVD $X_t = U \sigma(X_t) V^\top$

Gradient step $X_{t+\frac{1}{2}} = U \text{diag}(\nabla \psi(\sigma(X_t))) V^\top - \eta f'_t(X_t)$

Compute SVD $X_{t+\frac{1}{2}} = \tilde{U} \sigma(X_{t+\frac{1}{2}}) \tilde{V}^\top$

Shrinkage $X_{t+1} = \tilde{U} \text{diag} \left[ (\nabla \psi)^{-1} S_{\eta \lambda} \sigma(X_{t+\frac{1}{2}}) \right] \tilde{V}^\top$
Trace-norm Regularization Example

Proximal function:

\[ \psi(X) = \frac{1}{2} \|X\|_2^2 = \frac{1}{2} \|X\|_{Fr}^2 \]

Update:

\[ X_{t+\frac{1}{2}} = X_t - \eta f_t'(X_t) \quad (= U\Sigma_{t+\frac{1}{2}} V^\top) \]

Shrinkage:

\[ X_{t+1} = U \left[ \Sigma_{t+\frac{1}{2}} - \eta \lambda \right]_+ V^\top \]
Trace-norm Regularization Example

Proximal function:

$$\psi(X) = \frac{1}{2} \|X\|_2^2 = \frac{1}{2} \|X\|_{\text{Fr}}^2$$

Update:

$$X_{t+\frac{1}{2}} = X_t - \eta f'_t(X_t) \ (= U\Sigma_{t+\frac{1}{2}} V^\top)$$

Shrinkage:

$$X_{t+1} = U \left[ \Sigma_{t+\frac{1}{2}} - \eta \lambda \right]_+ V^\top$$
Proof ideas for trace-norm

**Idea:** Unitary invariance to reduce to vector case (Lewis 1995)

\[ \nabla \psi(X) = U \text{diag} [\nabla \psi(\sigma(X))] V^T \]

\[ \partial \|X\|_1 = U \text{diag}(\partial \|\sigma(X)\|_1) V^T \]

Simply reduce to vector case with \( \ell_1 \)-regularization
Conclusions and Related Work

- All derivations apply to Regularized Dual Averaging (Xiao 2009)

\[ x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\{ \eta \sum_{\tau=1}^{t} \langle g_{\tau}, x \rangle + \eta t \varphi(x) + \psi(x) \right\} \]

- Analysis of online convex programming for regularized objectives

- Unify several previous algorithms (projected gradient, mirror descent, forward-backward splitting)

- Derived algorithms for several regularization functions