Composite Objective Mirror Descent

Presentation by: Mojtaba Kadkhodaie

University of Minnesota

February 5, 2014

Reference

- J. Duchi, S. Shalev-Shwartz, Y. Singer, A. Tewari, "Composite Objective Mirror Descent," *Conference on Learning Theory (COLT), 2010*.

Agenda

- Introduction
- Composite Objective Mirror Descent (COMID)
- Regret Analysis
- Convergence Analysis
- Special Cases

Introduction

- Regularized Loss Minimization Problem

$$\min_{\mathbf{w}\in\Omega} \quad \frac{1}{n}\sum_{i=1}^{n} f_t(\mathbf{w}) + r(\mathbf{w}) \tag{1}$$

- $\Omega \subseteq \mathbb{R}^d$ closed convex set
- $f_t : \Omega \to \mathbb{R}$ convex loss function
- $r : \Omega \to \mathbb{R}$ convex regularization function

- **Examples:** Least Squares, LASSO, Ridge Regression, SVM,...

- 1^{st} order algorithm: Only sub-gradients are available
- Online Optimization: A single loss function f_t is accessed in each iteration.

Online Mirror Descent (OMD) Algorithm

- Convex Optimization Problem

$$\min_{\mathbf{w}\in\Omega} \frac{1}{n} \sum_{t=1}^{T} \phi_t(\mathbf{w}) \tag{2}$$

- Online Mirror Descent (OMD) Update Rule:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} \left\{ \eta \left\langle \phi_t'(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \right\rangle + B_\psi(\mathbf{w}, \mathbf{w}_t) \right\}$$

  - $\phi_t'(\mathbf{w}_t)$: Arbitrary sub-gradient of ϕ_t at $\mathbf{w}_t$
  - η: step-size (trade-off parameter)
  - $B_\psi(\mathbf{w}, \mathbf{w}_t)$: Bregman Divergence

$$B_\psi(\mathbf{w}, \mathbf{w}_t) = \psi(\mathbf{w}) - \psi(\mathbf{w}_t) - \langle \nabla\psi(\mathbf{w}), \mathbf{w} - \mathbf{w}_t \rangle \tag{3}$$

  - $\psi(\mathbf{w})$: differentiable and α-strongly convex w.r.t. a norm $\|\cdot\|$

$$B_\psi(\mathbf{w}, \mathbf{w}_t) \geq \frac{\alpha}{2} \|\mathbf{w} - \mathbf{w}_t\|^2$$

Online Mirror Descent (Cont.)

- **Example:** Assume $\Omega = \mathbb{R}^d$ and $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$.
  - OMD reduces to Online Gradient Descent (OGD) algorithm

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \left\{ \eta \left\langle \phi_t^{'}(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \right\rangle + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|_2^2 \right\}. \quad (4)$$

  - Optimality condition of (4) at $\mathbf{w}_{t+1}$ implies

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\, \phi_t^{'}(\mathbf{w}_t).$$

- OMD ignores the composite structure of the objective ϕ_t.

Composite Objective Mirror Descent (COMID) Algorithm

- Let $\phi_t = f_t + r$.
- COMID modifies the mirror descent update:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} \left\{ \eta \left\langle f_t^{'}(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \right\rangle + \eta\, r(\mathbf{w}) + B_\psi(\mathbf{w}, \mathbf{w}_t) \right\}. \tag{5}$$

- COMID does not linearize the regularization function $r(\mathbf{w})$.
- **Example:** Assume $\Omega = \mathbb{R}^d$ and $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$.
  - COMID is equivalent to the following two-step process:

1: Step One:

$$\tilde{\mathbf{w}}_{t+1} = \mathbf{w}_t - \eta\, f_t^{'}(\mathbf{w}_t)$$

2: Step Two:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \{\eta\, r(\mathbf{w}) + B_\psi(\mathbf{w}, \tilde{\mathbf{w}}_{t+1})\}$$

COMID: Two-Step Update

- Two-step update for general Bregman function ψ ($\Omega = \mathbb{R}^d$):
  - Step one:

$$\tilde{\mathbf{w}}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \eta \left\langle f_t^{'}(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t \right\rangle + B_\psi(\mathbf{w}, \mathbf{w}_t) \right\} \quad (6)$$

  - Step two:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left\{ \eta\, r(\mathbf{w}) + B_\psi(\mathbf{w}, \tilde{\mathbf{w}}_{t+1}) \right\}. \quad (7)$$

- Two-step Optimality Conditions:
  - Step one: $\eta\, f_t^{'}(\mathbf{w}_t) + \nabla\psi(\tilde{\mathbf{w}}_{t+1}) - \nabla\psi(\mathbf{w}_t) = 0$
  - Step two: $\eta\, r^{'}(\mathbf{w}_{t+1}) + \nabla\psi(\mathbf{w}_{t+1}) - \nabla\psi(\tilde{\mathbf{w}}_{t+1}) = 0$

Optimality Condition of COMID, Eq. (5):

$$\eta\, f_t^{'}(\mathbf{w}_t) + \eta\, r^{'}(\mathbf{w}_{t+1}) + \nabla\psi(\mathbf{w}_{t+1}) - \nabla\psi(\mathbf{w}_t) = 0 \quad (8)$$

Regret Bounds

- Regularized regret against a comparator $\mathbf{w}^*$

$$R(T) \doteq \sum_{t=1}^{T} (f_t(\mathbf{w}_t) + r(\mathbf{w}_t) - f_t(\mathbf{w}^*) - r(\mathbf{w}^*)) \tag{9}$$

Lemma 1: Let

- $\{\mathbf{w}_t\}$ be the sequence of COMID iterates with $\eta_t = \eta$,
- $B_\psi(\mathbf{w}, \mathbf{v}) \geq \frac{\alpha}{2}\|\mathbf{w} - \mathbf{v}\|^2$.

Then for any $\mathbf{w}^* \in \Omega$ we have

$$\eta\left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)\right) + \eta\left(r(\mathbf{w}_{t+1}) - r(\mathbf{w}^*)\right)$$
$$\leq B_\psi(\mathbf{w}^*, \mathbf{w}_t) - B_\psi(\mathbf{w}^*, \mathbf{w}_{t+1}) + \frac{\eta^2}{2\alpha}\|f_t^{'}(\mathbf{w}_t)\|_*^2.$$

Regret Bounds (Cont.)

By Lemma 1,

$$\eta \sum_{t=1}^{T}[f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*) + r(\mathbf{w}_{t+1}) - r(\mathbf{w}^*)]$$

$$\leq B_\psi(\mathbf{w}^*, \mathbf{w}_1) - B_\psi(\mathbf{w}^*, \mathbf{w}_{T+1}) + \frac{\eta^2}{2\alpha}\sum_{t=1}^{T}\|f_t^{'}(\mathbf{w}_t)\|_*^2.$$

Since $r(\cdot)$ and $B(\cdot,\cdot)$ are non-negative

$$\eta \sum_{t=1}^{T}[f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*) + r(\mathbf{w}_t) - r(\mathbf{w}^*)]$$

$$\leq B_\psi(\mathbf{w}^*, \mathbf{w}_1) + \frac{\eta^2}{2\alpha}\sum_{t=1}^{T}\|f_t^{'}(\mathbf{w}_t)\|_*^2 + r(\mathbf{w}_1)$$

Assume functions f_t are Lipschitz continuous $\|f_t^{'}(\mathbf{w}_t)\|_* \leq G_*$, then

$$R(T) \leq \frac{1}{\eta}B_\psi(\mathbf{w}^*, \mathbf{w}_1) + r(\mathbf{w}_1) + \frac{T\eta}{2\alpha}G_*^2$$

Regret Bounds: Theorem 1

Theorem 1: Let

- $\{\mathbf{w}_t\}$ be the sequence of COMID iterates,
- $B_\psi(\mathbf{w}, \mathbf{v}) \geq \frac{\alpha}{2}\|\mathbf{w} - \mathbf{v}\|^2$,
- Functions f_t be Lipschitz so that $\|f_t^{'}\|_* \leq G_*$
- $r(\mathbf{w}_1) = 0$.

Then for any $\mathbf{w}^* \in \Omega$ we have

$$R(T) \leq \frac{1}{\eta} B_\psi(\mathbf{w}^*, \mathbf{w}_1) + \frac{T\eta}{2\alpha} G_*^2. \tag{10}$$

In particular, by setting $\eta = \sqrt{2\alpha B_\psi(\mathbf{w}^*, \mathbf{w}_1)}/(G_*\sqrt{T})$ we get

$$R(T) \leq \sqrt{2TB_\psi(\mathbf{w}^*, \mathbf{w}_1)} G_*/\sqrt{\alpha}.$$

Proof of Lemma 1

By the convexity of f_t and r we have

$$\begin{aligned}
&\eta\left[f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*) + r(\mathbf{w}_{t+1}) - r(\mathbf{w}^*)\right] \\
&\quad \leq \eta\langle \mathbf{w}_t - \mathbf{w}^*, f_t^{'}(\mathbf{w}_t)\rangle + \eta\langle \mathbf{w}_t - \mathbf{w}^*, r^{'}(\mathbf{w}_{t+1})\rangle \\
&\quad = \langle \mathbf{w}^* - \mathbf{w}_{t+1}, \nabla\psi(\mathbf{w}_t) - \nabla\psi(\mathbf{w}_{t+1}) - \eta f_t^{'}(\mathbf{w}_t) - \eta\, r^{'}(\mathbf{w}_{t+1})\rangle \\
&\quad + \langle \mathbf{w}^* - \mathbf{w}_{t+1}, \nabla\psi(\mathbf{w}_{t+1}) - \nabla\psi(\mathbf{w}_t)\rangle + \eta\left\langle \mathbf{w}_t - \mathbf{w}_{t+1}, f_t^{'}(\mathbf{w}_t)\right\rangle
\end{aligned}$$

The purple term is non-positive, due to the optimality of $\mathbf{w}_{t+1}$.

$$\begin{aligned}
&\langle \mathbf{w}^* - \mathbf{w}_{t+1}, \nabla\psi(\mathbf{w}_{t+1}) - \nabla\psi(\mathbf{w}_t)\rangle \\
&\qquad = B_\psi(\mathbf{w}^*, \mathbf{w}_t) - B_\psi(\mathbf{w}^*, \mathbf{w}_{t+1}) - B_\psi(\mathbf{w}_{t+1}, \mathbf{w}_t) \\
&\qquad \leq B_\psi(\mathbf{w}^*, \mathbf{w}_t) - B_\psi(\mathbf{w}^*, \mathbf{w}_{t+1}) - \frac{\alpha}{2}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2
\end{aligned}$$

$$\begin{aligned}
\eta\left\langle \mathbf{w}_t - \mathbf{w}_{t+1}, f_t^{'}(\mathbf{w}_t)\right\rangle &= \eta\left\langle \sqrt{\frac{\alpha}{\eta}}(\mathbf{w}_t - \mathbf{w}_{t+1}), \sqrt{\frac{\eta}{\alpha}} f_t^{'}(\mathbf{w}_t)\right\rangle \\
&\leq \frac{\alpha}{2}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 + \frac{\eta^2}{2\alpha}\|f_t^{'}(\mathbf{w}_t)\|_*^2
\end{aligned}$$

Regret Bounds: Strongly Convex Problems

- Let $\phi_t = f_t + r$ be λ-strongly convex w.r.t. a function ψ:

$$\phi_t(\mathbf{v}) \geq \phi_t(\mathbf{w}) + \langle \phi_t^{'}(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \lambda B_\psi(\mathbf{v}, \mathbf{w}).$$

Lemma 2: Let

- $\{\mathbf{w}_t\}$ be the COMID iterates with step-size η_t,
- $r(\cdot)$ be λ-strongly convex w.r.t. a differentiable ψ,
- $B_\psi(\mathbf{w}, \mathbf{v}) \geq \frac{\alpha}{2}\|\mathbf{w} - \mathbf{v}\|^2$.

Then for any $\mathbf{w}^*$

$$\eta_t \left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)\right) + \eta_t \left(r(\mathbf{w}_{t+1}) - r(\mathbf{w}^*)\right)$$
$$\leq B_\psi(\mathbf{w}^*, \mathbf{w}_t) - B_\psi(\mathbf{w}^*, \mathbf{w}_{t+1}) + \frac{\eta_t^2}{2\alpha}\|f_t^{'}(\mathbf{w}_t)\|_*^2$$
$$-\lambda \eta_t B_\psi(\mathbf{w}^*, \mathbf{w}_{t+1}).$$

- The proof is very similar to Lemma 1.

Regret Bounds: Theorem 2

Theorem 2: Let

- $\{\mathbf{w}_t\}$ be sequence of the COMID iterates,
- $r(\cdot)$ be λ-strongly convex w.r.t. a differentiable ψ,
- $B_\psi(\mathbf{w}, \mathbf{v}) \geq \frac{\alpha}{2}\|\mathbf{w} - \mathbf{v}\|^2$,
- $\|f_t^{'}(\mathbf{w}_t)\|_* \leq G_*$
- $r(\mathbf{w}_1) = 0$
- $\eta_t = \frac{1}{\lambda t}$.

Then for any $\mathbf{w}^* \in \Omega$

$$R(T) \leq \lambda B_\psi(\mathbf{w}^*, \mathbf{w}_1) + \frac{G_*^2}{\lambda\alpha}(\log T + 1) = \mathcal{O}\left(\frac{G_*^2}{\lambda\,\alpha}\log T\right)$$

Example: $r(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ (Ridge Regression)

Stochastic COMID

- A random subset of loss functions $\{f_t\}$ is available each time.
- Define $f(\mathbf{w}) = \mathbb{E}f(\mathbf{w}; Z) = \int f(\mathbf{w}; z)dPz$.
- At every step t, an independent $Z_t \sim P$ gives

$$f_t(\mathbf{w}_t) = f(\mathbf{w}_t; Z_t)$$

which is an unbiased estimate of $f(\mathbf{w}_t)$ and also gives

$$f_t^{'}(\mathbf{w}_t) = f^{'}(\mathbf{w}_t; Z_t)$$

which is unbiased estimate of an arbitrary $f^{'}(\mathbf{w}_t) \in \partial f(\mathbf{w}_t)$.

Stochastic Convergence Analysis

Theorem 3: Let

- $\{\mathbf{w}_t\}$ be sequence of the COMID iterates,
- $B_\psi(\mathbf{w}^*, \mathbf{w}_t) \leq D^2, \quad \forall t,$
- $\|f_t^{'}(\mathbf{w}_t)\|_* \leq G_*$
- $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$
- $\eta_t = \frac{D}{G_*\sqrt{\alpha t}}$

Then, with probability at least $1-\delta$

$$f(\bar{\mathbf{w}}_T)+r(\bar{\mathbf{w}}_T) \leq f(\mathbf{w}^*)+r(\mathbf{w}^*)+\frac{DG_*}{\sqrt{\alpha T}}\left(1+4\sqrt{\log\frac{1}{\delta}}\right).$$

If Ω is compact, assumptions (2) and (3) are satisfied.

Special Cases: p-norm Divergences

- Already studied the case where $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ (slide 6).
- Now consider the case where $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_p^2, \;\; p \in (1, 2]$.
- **Fact:** $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_p^2$ is $(p-1)$-strongly convex w.r.t. the ℓ_p-norm over $\mathbb{R}^d$.

Corollary 1: Let

- $p = 1 + 1/\log d$
- $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_p^2$,
- $\|f_t'(\mathbf{w})\|_q \leq G_q$, when $q = \log d + 1$,
- $\eta = \frac{\|\mathbf{w}^*\|_p}{G_q}\sqrt{\frac{1}{T \log d}}$.

Then the COMID regret satisfies

$$R(T) \leq \|\mathbf{w}^*\|_p G_q \sqrt{T \log d}.$$

- Interesting case is when d is large.

Special Cases: p-norm Divergences

- Let $\psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_p^2$ and $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$.
- The COMID update is
  1: Step One:

$$\nabla\psi(\tilde{\mathbf{w}}_{t+1}) = \nabla\psi(\mathbf{w}_t) - \eta\, f_t^{'}(\mathbf{w}_t)$$

  2: Step Two:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \{\eta\,\lambda\,\|\mathbf{w}\|_1 + B_\psi(\mathbf{w}, \tilde{\mathbf{w}}_{t+1})\}$$

- Step two is equivalent to

$$\nabla\psi(\mathbf{w}_{t+1}) = \mathcal{S}_{\eta\,\lambda}(\nabla\psi(\tilde{\mathbf{w}}_{t+1})),$$

where $\mathcal{S}_\tau$ is the shrinkage/thresholding operator

$$[\mathcal{S}_\tau(\mathbf{x})]_j = \mathsf{sign}(x_j)[|x_j| - \tau]_+ \ .$$

Matrix Composite Mirror Descent

- Let $\mathbf{W} \in \Omega = \mathbb{R}^{d_1 \times d_2}$.
- Let Bregman functions be the Schatten p-norms, i.e.

$$\|\mathbf{W}\|_p = \|\sigma(\mathbf{W})\|_p$$

- **Fact:** $\psi(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\|_p^2$ is $(p-1)$-strongly convex w.r.t. the Schatten p-norm over $\mathbb{R}^{d_1 \times d_2}$.
- Let $r(\mathbf{W}) = \lambda\|\mathbf{W}\|_1$ (Nuclear Norm).
- The COMID update is

$$\mathbf{W}_{t+1} = \arg\min_{\mathbf{W}\in\Omega} \left\{ \eta \left\langle f_t^{'}(\mathbf{W}_t), \mathbf{W} - \mathbf{W}_t \right\rangle + \eta\,\lambda\,\|\mathbf{W}\|_1 + B_\psi(\mathbf{W}, \mathbf{W}_t) \right\}.$$

Matrix Composite Mirror Descent (Cont.)

- Let $\psi(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\|_2^2$. Then the COMID update is equivalent to

$$\text{Compute SVD: } \mathbf{W}_t = \mathbf{U}_t \,\text{diag}(\sigma(\mathbf{W}_t))\, \mathbf{V}_t^T$$

$$\text{Gradient Step: } \mathbf{Z}_t = \mathbf{W}_t - \eta f_t^{'}(\mathbf{W}_t)$$

$$\text{Compute SVD: } \mathbf{Z}_t = \tilde{\mathbf{U}}_t \text{diag}(\sigma(\mathbf{Z}_t))\tilde{\mathbf{V}}_t^T$$

$$\text{Splitting Update: } \mathbf{W}_{t+1} = \tilde{\mathbf{U}}_t \,\text{diag}(\mathcal{S}_{\eta\lambda}(\sigma(\mathbf{Z}_t)))\, \tilde{\mathbf{V}}_t^T$$

Corollary 2: Let

- $\psi(\mathbf{W}) = \frac{1}{2}\|\mathbf{W}\|_2^2$,
- $\|f_t^{'}(\mathbf{W})\|_2 \leq G_2$,

Then the COMID regret satisfies

$$R(T) \leq \|\mathbf{W}^*\|_2 \, G_2 \sqrt{T}.$$

Thank You!