Reading Group for Online Learning

Week 4: Adaptive Online Convex Optimization

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

October 10, 2017
Outline

Online Convex Optimization

Adaptive Online Mirror Descent (OMD)
  From OMD to Adaptive OMD
  Basic Analysis for Adaptive OMD
  Isotropically Adaptive Online Gradient Descent
  Diagonal AdaGrad
  Online Newton Step

Regret with Gradual Variations


Online Convex Optimization

• Player chooses a vector $x_t$ from a convex compact set $\mathcal{X} \subseteq \mathbb{R}^d$.
• Player observes a convex loss function $f_t : \mathcal{X} \rightarrow \mathbb{R}^d$, suffers the loss $f_t(x_t)$, and receives its subgradient $\nabla f_t(x_t)$ as feedback.

Example

• Regression with square loss: $f_t(x) = (x_t^\top a_t - b_t)^2$
• Classification with hinge loss: $f_t(x) = (1 - b_t x_t^\top a_t)_+, b_t \in \{-1, 1\}$.

Regret

The goal is to minimize the cumulative loss (i.e., regret):

$$R_T = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} f_t(x).$$
Online Mirror Descent (OMD)

- Let \( \psi : \mathcal{X} \rightarrow \mathbb{R} \) be differentiable and 1-strongly convex with respect to a norm \( \| \cdot \| \). Such a function is called as a mirror map (w.r.t. \( \| \cdot \| \)).
- The Bregman divergence of \( x, y \in \mathcal{X} \) with respect to \( \psi \) is given by

\[
B^\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle.
\]

### Online Mirror Descent

1. Choose \( x_1 \in \mathbb{R}^d \), a mirror map \( \psi \), and \( \{\gamma_t > 0\}_{t=1}^T \).
2. In round \( t \):
   2a. Receive \( f_t \), compute \( g_t = \nabla f_t(x_t) \).
   2b. \( x_{t+1} = \arg \min_{x \in \mathcal{X}} \{\langle g_t, x \rangle + \gamma_t^{-1} B^\psi(x, x_t)\} \).

- If \( \psi = \frac{1}{2} \| \cdot \|^2_2 \), then \( B^\psi(x, y) = \frac{1}{2} \| x - y \|^2_2 \). The algorithm is online gradient descent.

### Convergence

Let \( \gamma_t = \gamma / \sqrt{t} \) for some \( \gamma > 0 \).

\[
R_T \leq \left( \frac{1}{\gamma} \max_{x \in \mathcal{X}, t \leq T} B^\psi(x, x_t) + \frac{\gamma}{2} \max_{t \leq T} \| g_t \|_*^2 \right) \sqrt{T}
\]

- \( \| \cdot \|_* \) is the dual norm of \( \| \cdot \| \), i.e., \( \| x \|_* = \max_{\| y \| \leq 1} \langle x, y \rangle \).
Outline

Online Convex Optimization

Adaptive Online Mirror Descent (OMD)
   From OMD to Adaptive OMD
   Basic Analysis for Adaptive OMD
   Isotropically Adaptive Online Gradient Descent
   Diagonal AdaGrad
   Online Newton Step

Regret with Gradual Variations
Generalized OMD

Online Mirror Descent

1. Choose $x_1 \in \mathbb{R}^d$, a mirror map $\psi$ and $\{\gamma_t > 0\}^T_{t=1}$.
2. In round $t$:
   2a. Receive $f_t$, compute $g_t = \nabla f_t(x_t)$.
   2b. $x_{t+1} = \arg\min_{x \in \mathcal{X}} \{\langle g_t, x \rangle + \gamma_t^{-1} B^\psi(x, x_t)\}$.

- Fixed mirror map $\psi$, (possibly) time-varying step-size $\gamma_t$.

Generalized Online Mirror Descent

1. Choose $x_1 \in \mathbb{R}^d$.
2. In round $t$:
   2a. Receive $f_t$, compute $g_t = \nabla f_t(x_t)$, and choose a mirror map $\psi_t$,
   2b. $x_{t+1} = \arg\min_{x \in \mathcal{X}} \{g_t, x\} + B^\psi_t(x, x_t)\}$.

- $\psi_t$ could be time-varying.
- When $\psi_t = \gamma_t^{-1}\psi$, it reduces to the OMD with time-varying step-size.
Adaptive Online Mirror Descent

1. Choose $x_1 \in \mathbb{R}^d$.
2. In round $t$:
   2a. Receive $f_t$, compute $g_t = \nabla f_t(x_t)$, and choose a mirror map $\psi_t$,
   2b. $x_{t+1} = \arg \min_{x \in X} \{\langle g_t, x \rangle + B_{\psi_t}(x, x_t)\}$.

- $\psi_t$ is an (adaptive) regularization function which may depend on $\{(x_s, g_s)\}_{s \leq t}$. 
Adaptive Strategies for OMD

In what follows, we will study the following three different choices of $\psi_t$.

**Isotropically Adaptive OGD**

Let $\max_{x,y \in X} \|x - y\|_2 \leq b$ and $\psi_t = \| \cdot \|_2 \sqrt{\sum_{s \leq t} \|g_s\|_2^2} / b$. Then

$$R_T \leq 1.5b \sqrt{\sum_{s=1}^{T} \|g_s\|_2^2}$$

**Diagonal AdaGrad [3]**

Let $G_t = \epsilon I + \sum_{s \leq t} g_s g_s^\top$ and $S_t = \text{diag}(G_t)$. Let $\max_{x,y \in X} \|x - y\|_\infty \leq b_\infty$.

Choose $\psi_t = \| \cdot \|_2^2 / b_\infty$. Then $R_T \leq 1.5b_\infty \sum_{i=1}^{d} \sqrt{\sum_{s=1}^{T} \|g_s(i)\|_2^2}$

**Online Newton Step [4]**

Let $f_t$ be $\beta$-exp-concave. Choose $\psi_t = \beta \| \cdot \|^2_{G_t} / 2$. Then $R_T \leq \log T$. 
Some Basic Notations and Lemmas

**Strong Convexity**

A function \( \psi : \mathcal{X} \to \mathbb{R} \) is \( \mu \)-strongly convex w.r.t. a norm \( \| \cdot \| \) if for all \( x, y \in \mathcal{X} \),

\[
\psi(x) - \psi(y) \geq \nabla \psi(y)^\top (x - y) + \frac{\mu}{2} \| x - y \|^2.
\]

**Example**

- **Euclidean norm**: \( \psi = \frac{1}{2} \| \cdot \|_2^2 \) is 1-strongly convex w.r.t. \( \| \cdot \|_2 \).
- **Power norm**: for a given positive definite matrix \( H \in \mathbb{R}^{d \times d} \), \( \psi(x) = \frac{1}{2} x^\top H x \) is 1-strongly convex w.r.t. the norm \( \| \cdot \|_H = \sqrt{x^\top H x} \).

**Weierstrass Extreme Value Theorem**

Every continuous function on a compact set attains its extreme values on that set.

**First-order Condition for Convex Minimization**

Let \( f(x) : \mathcal{X} \to \mathbb{R} \) be differentiable and convex. Let \( \hat{x} = \arg \min_{x \in \mathcal{X}} f(x) \). Then

\[
\langle \nabla f(\hat{x}), y - \hat{x} \rangle \geq 0, \quad \forall y \in \mathcal{X}.
\]
**Bregman divergences**

**Definition (Bregman divergence)**

Let \( \psi : \mathcal{X} \rightarrow \mathbb{R} \) be a continuously-differentiable and 1-strictly convex function (w.r.t. \( \| \cdot \| \)) defined on a compact convex set \( \mathcal{X} \). The Bregman divergence \( B^{\psi} \) associated with \( \psi \) for points \( x \) and \( y \) is:

\[
B^{\psi}(x, y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle
\]

- \( \psi(\cdot) \) is referred to as the Bregman function, or the mirror map.
- The Bregman divergence satisfies the following properties:
  (a) by the strong convexity of \( \psi \),
  \[
  B^{\psi}(x, y) \geq 1/2 \| x - y \|^2
  \]
  (b) \[
  \frac{\partial B^{\psi}(x, y)}{\partial x} = \nabla \psi(x) - \nabla \psi(y)
  \]
  (c) For all \( x, y, z \in \mathcal{X} \),
  \[
  B^{\psi}(x, y) = B^{\psi}(x, z) + B^{\psi}(z, y) + \langle (x - z), \nabla \psi(y) - \nabla \psi(z) \rangle
  \]
  (d) For all \( x, y \in \mathcal{X} \),
  \[
  B^{\psi}(x, y) + B^{\psi}(y, x) = \langle (x - y), \nabla \psi(x) - \nabla \psi(y) \rangle
  \]
- \( B^{\psi}(x, y) \neq B^{\psi}(y, x) \) in general.
*Bregman divergences*

- The Bregman divergence is the **vertical distance** at \( x \) between \( \psi \) and the **tangent** of \( \psi \) at \( y \), see figure below

- The Bregman divergence measures the **strictness of convexity** of \( \psi(\cdot) \).

**Example**

- Let \( \| \cdot \| = \frac{1}{\sqrt{\gamma}} \| \cdot \|_2 \), and \( \psi = \frac{1}{2\gamma} \| \cdot \|_2^2 \). Then \( B^\psi(x, y) = \frac{1}{2\gamma} \| x - y \|_2^2 \).
- Let \( A \in \mathbb{R}^{d \times d} \) be positive definite. \( \| x \| = \sqrt{x^\top A x} \). Then \( \psi(x) = \frac{1}{2} x^\top A x \) is strongly-convex and \( B^\psi(x, y) = \frac{1}{2} \| x - y \|_A^2 \).
Equivalent Form for Mirror Descent Step

- Recall that the mirror descent step at $t$ round is as follows:

$$x_{t+1} = \arg \min_{x \in \mathcal{X}} \{ \langle g_t, x \rangle + B_{\psi_t}(x, x_t) \}. \quad (1)$$

**Lemma (A)**

*Equation (1) is equivalent to*\(^{1}\)

$$\begin{cases}
\nabla \psi_t(y_{t+1}) = \nabla \psi_t(x_t) - g_t \\
x_{t+1} = \arg \min_{x \in \mathcal{X}} B_{\psi_t}(x, y_{t+1}).
\end{cases}$$

**Proof.** The proof is straightforward.

\(^{1}\text{Here, we assume the existence of } y_{t+1} \text{ such that it satisfies the first equation. The existence of } y_{t+1} \text{ is always guaranteed if there exists a domain } \mathcal{D} \text{ such that } \mathcal{X} \subseteq \mathcal{D} \text{ and } \nabla \psi_t(\mathcal{D}) = \mathbb{R}^d.\)
Basic Property for Mirror Descent Step

\[
\begin{align*}
\psi_t(y_{t+1}) &= \nabla \psi_t(x_t) - g_t \\
x_{t+1} &= \arg\min_{x \in \mathcal{X}} B^{\psi_t}(x, y_{t+1}).
\end{align*}
\]

(1)

Lemma A

Let \(x_{t+1}\) be given by (1). Then for any \(x \in \mathcal{X}\),

\[
\langle g_t, x_{t+1} - x \rangle \leq B^{\psi_t}(x, x_t) - B^{\psi_t}(x, x_{t+1}) - B^{\psi_t}(x_{t+1}, x_t).
\]

Proof. According to the first equality of (1),

\[
\langle g_t, x_{t+1} - x \rangle = \langle \nabla \psi_t(x_t) - \nabla \psi_t(y_{t+1}), x_{t+1} - x \rangle \\
= \langle \nabla \psi_t(x_t) - \nabla \psi_t(x_{t+1}), x_{t+1} - x \rangle + \langle \nabla \psi_t(x_{t+1}) - \nabla \psi_t(y_{t+1}), x_{t+1} - x \rangle.
\]

Since \(x_{t+1}\) is the minimizer, according to the first order optimality condition,

\[
0 \geq \langle \nabla B^{\psi_t}(x, y_{t+1})|_{x=x_{t+1}}, x_{t+1} - x \rangle = \langle \nabla \psi_t(x_{t+1}) - \nabla \psi_t(y_{t+1}), x_{t+1} - x \rangle.
\]

Thus,

\[
\langle g_t, x_{t+1} - x \rangle \leq \langle \nabla \psi_t(x_t) - \nabla \psi_t(x_{t+1}), x_{t+1} - x \rangle \\
= B^{\psi_t}(x, x_t) - B^{\psi_t}(x, x_{t+1}) - B^{\psi_t}(x_{t+1}, x_t).
\]
\[
\begin{align*}
\nabla \psi_t(y_{t+1}) &= \nabla \psi_t(x_t) - g_t \\
x_{t+1} &= \arg\min_{x \in \mathcal{X}} B_{\psi_t}(x, y_{t+1}).
\end{align*}
\]

**Lemma A**

Let \( x_{t+1} \) be given by (1). Then for any \( x \in \mathcal{X} \),

\[
\langle g_t, x_{t+1} - x \rangle \leq B_{\psi_t}(x, x_t) - B_{\psi_t}(x, x_{t+1}) - B_{\psi_t}(x_{t+1}, x_t).
\]

By Cauchy-Swartz inequality,

\[
\begin{align*}
\langle g_t, x_t - x \rangle &= \langle g_t, x_{t+1} - x \rangle + \langle g_t, x_t - x_{t+1} \rangle \\
&\leq \langle g_t, x_{t+1} - x \rangle + \|g_t\|_{t,*} \|x_t - x_{t+1}\|_t \\
&\leq \langle g_t, x_{t+1} - x \rangle + \frac{1}{2} \|g_t\|_{t,*}^2 + \frac{1}{2} \|x_t - x_{t+1}\|_t^2 \\
&\leq \langle g_t, x_{t+1} - x \rangle + \frac{1}{2} \|g_t\|_{t,*}^2 + \frac{1}{2} B_{\psi_t}(x_{t+1}, x_t),
\end{align*}
\]

where for the last inequality, we used the strongly convexity of \( \psi_t \) w.r.t. the norm \( \| \cdot \|_t \). Combining with Lemma a, we thus get

\[
\langle g_t, x_t - x \rangle \leq B_{\psi_t}(x, x_t) - B_{\psi_t}(x, x_{t+1}) + \frac{1}{2} \|g_t\|_{t,*}^2,
\]
So far, we have proved that

$$\langle g_t, x_t - x \rangle \leq B^{\psi t}(x, x_t) - B^{\psi t}(x, x_{t+1}) + \frac{1}{2} \|g_t\|_{t,*}^2,$$

Writing $B^{\psi t}(x, x_t) - B^{\psi t}(x, x_{t+1})$ as

$$\left( B^{\psi t}(x, x_t) - B^{\psi t+1}(x, x_{t+1}) \right) + \left( B^{\psi t+1}(x, x_{t+1}) - B^{\psi t}(x, x_{t+1}) \right)$$

and summing up over $t = 1, \ldots, T$, one can easily prove that

$$\sum_{t=1}^{T} \langle g_t, x_t - x \rangle \leq B^{\psi 1}(x, x_1) - B^{\psi T}(x, x_{T+1})$$

$$+ \sum_{t=1}^{T-1} \left( B^{\psi t+1}(x, x_{t+1}) - B^{\psi t}(x, x_{t+1}) \right) + \frac{1}{2} \|g_t\|_{t,*}^2,$$

Noting that $B^{\psi T}(x, x_{T+1}) \geq 0$, we thus prove the following result.
**Basic Property of Mirror Descent**

**Proposition A**

Let $\psi_t$ be a mirror map (w.r.t. $\| \cdot \|_t$) and $\{x_t\}$ be defined by

$$\nabla \psi_t(y_{t+1}) = \nabla \psi_t(x_t) - g_t$$

$$x_{t+1} = \arg \min_{x \in X} B^\psi_t(x, y_{t+1}).$$

Then

$$\sum_{t=1}^{T} \langle g_t, x_t - x \rangle \leq B^\psi_1(x, x_1) + \sum_{t=2}^{T} B^\psi_t(x, x_t) - B^\psi_{t-1}(x, x_t) + \frac{1}{2} \sum_{t=1}^{T} \|g_t\|_t^2.$$
Isotropically Adaptive Online Gradient Descent (Isotropic Adagrad)

Adaptive Online Gradient Descent

1. Choose $x_1 \in \mathbb{R}^d$ and $\gamma > 0$.
2. In round $t$:
   
   2a. Compute $g_t = \nabla f_t(x_t)$ and $\gamma_t = \gamma / \sqrt{\sum_{s \leq t} \|g_s\|^2}$,
   
   2b. $x_{t+1} = \arg \min_{x \in X} \|x - (x_t - \gamma_t g_t)\|_2$.

- w.l.o.g. we assume that $\|g_1\|_2 > 0$.

Convergence

Let $\max_{x, y \in X} \|x - y\|_2 \leq b$. Then

$$R_T \leq (\gamma^{-1} b^2 / 2 + \gamma) \sqrt{\sum_{s=1}^{T} \|g_s\|^2}$$

- Adaptive learning rate.
Proof.

Let $\psi_t = \frac{1}{2\gamma_t} \| \cdot \|^2_2$. It is easy to see that $\psi$ is strongly convex with respect to $\| \cdot \|_t = \frac{1}{\sqrt{\gamma_t}} \| \cdot \|_2$. Moreover, $B^{\psi_t}(x, y) = \frac{1}{2\gamma_t} \| x - y \|^2_2$, and $\| \cdot \|_{t,*} = \sqrt{\gamma_t} \| \cdot \|_2$. A direct computation shows that, the iteration step in Adaptive OGD is equivalent to the iteration step in Proposition A. Thus, we can apply Proposition A to get

$$
\sum_{t=1}^T \langle g_t, x_t - x \rangle \leq \frac{1}{2\gamma_1} \| x - x_1 \|^2 + \sum_{t=2}^T \| x - x_t \|^2_2 \left( \frac{1}{2\gamma_t} - \frac{1}{2\gamma_{t-1}} \right) + \frac{1}{2} \sum_{t=1}^T \gamma_t \| g_t \|^2_2
$$

$$
\leq \frac{b^2}{2\gamma_T} + \frac{1}{2} \sum_{t=1}^T \gamma_t \| g_t \|^2_2 = \frac{b^2}{2\gamma} \sqrt{\sum_{t \leq T} \| g_t \|^2_2} + \frac{\gamma}{2} \sum_{t=1}^T \| g_t \|^2_2 / \sqrt{\sum_{s \leq t} \| g_s \|^2_2}.
$$

Using the basic inequality $\sum_{t=1}^T a_t (\sum_{s \leq t} a_s)^{-1/2} \leq 2 \sqrt{\sum_{t=1}^T a_t}$ for non-negative $a_t$ and $a_1 > 0$, and the convexity of $f_t$ which implies

$$
f_t(x_t) - f(x) \leq \langle g_t, x_t - x \rangle,
$$

one can prove the desired results.
Proof

**Lemma**

For non-negative $a_t$ and $a_1 > 0$,

$$
\sum_{t=1}^{T} \frac{a_t}{\sqrt{\sum_{s\leq t} a_s}} \leq 2 \sqrt{\sum_{t=1}^{T} a_t}
$$

**Proof** Let $L_t = \sum_{s\leq t} a_s$. Using the mean value theorem, one can prove that

$$
\frac{1}{2} x \leq 1 - \sqrt{1 - x}, \quad \forall x \in [0, 1].
$$

Applying the above inequality with $x = a_t/L_t$,

$$
\frac{1}{2} \frac{a_t}{L_t} \leq 1 - \sqrt{1 - \frac{a_t}{L_t}} = 1 - \sqrt{\frac{L_{t-1}}{L_t}},
$$

which leads to

$$
\frac{1}{2} \frac{a_t}{\sqrt{L_t}} \leq \sqrt{L_t} - \sqrt{L_{t-1}}.
$$

Summing up over $t = 1, \cdots T$, one can prove the desired result.
Diagonal AdaGrad: Motivation

• Recall that the online gradient descent is given by

\[
x_{t+1} = \arg\min_{x \in \mathcal{X}} \|x - (x_t - \gamma_t g_t)\|_2^2.
\]

For the special case \(\mathcal{X} \in \mathbb{R}^d\), it can be rewritten as,

\[
x_{t+1}(i) = x_t(i) - \gamma_t g_t(i).
\]

*All feature dimension share the same learning rate.*

• Practical examples often have high-dimensional feature spaces.
  – Many features are irrelevant
  – Rare features are often very informative.
Diagonal AdaGrad: Motivation

Why adapt to geometry?

<table>
<thead>
<tr>
<th>$y_t$</th>
<th>$\phi_{t,1}$</th>
<th>$\phi_{t,2}$</th>
<th>$\phi_{t,3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>-.5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>-1</td>
<td>-.5</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

1. Frequent, irrelevant
2. Infrequent, predictive
3. Infrequent, predictive

Examples from Duchi et al. ISMP 2012 slides
Diagonal AdaGrad: Motivation

- Standard online gradient descent:

\[
x_{t+1} = \arg \min_{x \in \mathcal{X}} \| x - (x_t - \gamma g_t) \|_2^2.
\]

It corresponds to OMD with \( \psi_t = \frac{1}{2\gamma} \| \cdot \|_2^2 \). Its regret bound:

\[
R_T \leq \frac{1}{2\gamma} \max_{x \in \mathcal{X}} \| x - x_1 \|_2^2 + \frac{\gamma}{2} \sum_{t=1}^{T} \left\| g_t \right\|_2^2.
\]

- What if we consider \( \psi_t = \frac{1}{2\gamma} \| \cdot \|_H^2 \)?

\[
x_{t+1} = \arg \min_{x \in \mathcal{X}} \| x - (x_t - \gamma H^{-1} g_t) \|_H^2
\]

The regret bound becomes:

\[
R_T \leq \frac{1}{2\gamma} \max_{x \in \mathcal{X}} \| x - x_1 \|_2^2 + \frac{\gamma}{2} \sum_{t=1}^{T} \left\| g_t \right\|_H^{-1}^2.
\]

*What H to choose?*
Diagonal AdaGrad: Motivation

\[ x_{t+1} = \arg \min_{x \in \mathcal{X}} \| x - (x_t - \gamma H^{-1} g_t) \|_H^2 \]

- The regret bound:

\[ R_T \leq \frac{1}{2\gamma} \max_{x \in \mathcal{X}} \| x - x_1 \|_2^2 + \frac{\gamma}{2} \sum_{t=1}^T \| g_t \|_{H^{-1}}^2. \]

- Choosing \( H \) to minimize

\[ \sum_{t=1}^T \| g_t \|_{H^{-1}}^2 = \text{Tr} \left( H^{-1} \sum_{t=1}^T g_t g_t^\top \right), \]

subject to \( H \succeq 0 \) and \( \text{Tr} (H) \leq C \), the solution is [3]

\[ H = c \left( \sum_{t \leq T} g_t g_t^\top \right)^{1/2} \]
Diagonal AdaGrad: Motivation

\[ x_{t+1} = \arg \min_{x \in \mathcal{X}} \| x - (x_t - \gamma H^{-1} g_t) \|_H^2 \]

- The regret bound:

\[ R_T \leq \frac{1}{2\gamma} \max_{x \in \mathcal{X}} \| x - x_1 \|_2^2 + \frac{\gamma}{2} \sum_{t=1}^{T} \| g_t \|^2_{H^{-1}}. \]

- Choosing \( H \) to minimize

\[ \sum_{t=1}^{T} \| g_t \|^2_{H^{-1}} = \text{Tr} \left( H^{-1} \sum_{t=1}^{T} g_t g_t^\top \right), \]

subject to \( H \succeq 0 \) and \( \text{Tr} \left( H \right) \leq C \), the solution is [3]

\[ H = c \left( \sum_{t \leq T} g_t g_t^\top \right)^{1/2} \]

- At time \( t \), replaced \( H \) by

\[ H_t = \left( \sum_{s \leq t} g_s g_s^\top \right)^{1/2} \]

and to reduce the computational cost, replace \( H_t \) by \( \text{diag}(H_t) \)
Diagonal AdaGrad

### AdaGrad

1. Choose $x_1 \in \mathbb{R}^d$ and $\gamma > 0$.
2. In round $t$:
   2a. Compute $g_t = \nabla f_t(x_t)$,
   2b. Update $A_t = [A_{t-1} \ g_t]$, $s_t(i) = \|A_t(i,:)\|^2_2$, $S_t = (\epsilon I + \text{diag}(s_t))^{1/2}$.
   2c. $x_{t+1} = \arg \min_{x \in \mathcal{X}} \|x - (x_t - \gamma S_t^{-1} g_t)\|^2_{S_t}$.

### Simple cases

Let $\mathcal{X} = \mathbb{R}^d$. Then the update step is, for each $i \in \{1, \cdots, d\}$,

$$x_{t+1}(i) = x_t(i) - \frac{\gamma}{\sqrt{\epsilon + \sum_{s \leq t} |g_s(i)|^2}} g_t(i)$$

- Each feature dimension has its own learning rate. The learning rate is adapted with $t$ and takes geometry of the past observations into account.
- When $|g_s(i)|^2 \approx 0$ for most $s$, then $\sum_{s \leq t} |g_t(i)|^2 \approx 0$. The learning rate in $i$-th feature tends to be larger, meaning that the $i$-th feature is important.
Diagonal AdaGrad: Convergence

**AdaGrad**

1. Choose $x_1 \in \mathbb{R}^d$ and $\gamma > 0$.
2. In round $t$:
   2a. Compute $g_t = \nabla f_t(x_t)$,
   2b. Update $A_t = [A_{t-1} \ g_t]$, $s_t(i) = \|A_t(i,:)\|_2^2$, $S_t = (\epsilon I + \text{diag}(s_t))^{1/2}$.
   2c. $x_{t+1} = \arg \min_{x \in \mathcal{X}} \|x - (x_t - \gamma S_t^{-1} g_t)\|_S^2$.

**Convergence**

Let $\max_{x,y \in \mathcal{X}} \|x - y\|_\infty \leq b_\infty$. Then

$$R_T \leq (b_\infty^2 / (2\gamma) + \gamma) \text{Tr} (S_T)$$

- Letting $\epsilon \to 0$, $\text{Tr} (S_T) \to \text{Tr} (\text{diag}(s_t)) = \sum_{i=1}^d \sqrt{\sum_{t=1}^T |g_t(i)|^2}$
AdaGrad Theoretical Examples

- Expect to perform well when the gradient vectors are sparse.

- SVM example. Let \( f_t(x) = (1 - b_t a_t^\top x) \) with \( a_t \in \{-1, 1, 0\}^d \). If \( a_t(i) \neq 0 \) with probability \( \propto i^{-\alpha} \) where \( \alpha > 1 \). Then
  \[
  \mathbb{E} R_T(x) \lesssim \sqrt{T \max(\log d, d^{1-\alpha/2})}.
  \]

- Previously regret bound:
  \[
  \mathbb{E} R_T(x) \lesssim \sqrt{T d}.
  \]
Diagonal AdaGrad: Theoretical Analysis

- Let \( \psi_t(\cdot) = \frac{1}{2\gamma} \| \cdot \|^2_{S_t} \). Then \( \psi \) is strongly convex w.r.t. \( \| \cdot \|_t = \frac{1}{\sqrt{\gamma}} \| \cdot \|_{S_t} \).

Moreover, \( B^{\psi_t}(x, y) = \frac{1}{2\gamma} \| x - y \|^2_{S_t} \), and \( \| \cdot \|_{t,*} = \sqrt{\gamma} \| \cdot \|_{S_t^{-1}} \).

- By a direct calculation, one can easily show that Diagonal AdaGrad can be written as

\[
\begin{aligned}
\nabla \psi_t(y_{t+1}) &= \nabla \psi_t(x_t) - g_t \\
x_{t+1} &= \arg \min_{x \in X} B^{\psi_t}(x, y_{t+1}).
\end{aligned}
\]

- Applying Proposition A,

\[
\sum_{t=1}^{T} \langle g_t, x_t - x \rangle \leq \frac{1}{2\gamma} \| x - x_1 \|^2_{S_1} + \frac{1}{2\gamma} \sum_{t=2}^{T} \| x - x_t \|^2_{S_t} - \| x - x_t \|^2_{S_t^{-1}} + \frac{\gamma}{2} \sum_{t=1}^{T} \| g_t \|^2_{S_t^{-1}}
\]

In the following, we will estimate the above three terms separately.

- \[ \frac{1}{2\gamma} \| x - x_1 \|^2_{S_1} = \frac{1}{2\gamma} (x - x_1)^\top S_1 (x - x_1) \leq \frac{\| x - x_1 \|^2_\infty}{2\gamma} \text{Tr} (S_1) \leq \frac{b_\infty^2}{2\gamma} \text{Tr} (S_1) \]
\[
\frac{1}{2\gamma} \sum_{t=2}^{T} \| x - x_t \|^2_{S_t} - \| x - x_t \|^2_{S_{t-1}} = \frac{1}{2\gamma} \sum_{t=2}^{T} (x - x_t)^\top (S_t - S_{t-1})(x - x_t)
\]
\[
\leq \frac{1}{2\gamma} \sum_{t=2}^{T} \| x - x_t \|^2_{\infty} \text{Tr} (S_t - S_{t-1}) \leq \frac{b^2_{\infty}}{2\gamma} \text{Tr} (S_T - S_1).
\]

\[
\sum_{t=1}^{T} \frac{\| g_t \|^2_{S_t^{-1}}}{S_t^{-1}} = \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{|g_t(i)|^2}{s_t(i)} = \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{|g_t(i)|^2}{\sqrt{\epsilon I + \sum_{j=1}^{t} |g_j(i)|}}
\]
\[
\leq 2 \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} |g_t(i)|^2},
\]

where for the last inequality we used the basic inequality
\[
2 \sum_{t=1}^{T} a_t (\sum_{s=1}^{t} a_s)^{-1/2} \leq (\sum_{t=1}^{T} a_t)^{1/2}.
\]

Combining the above estimates and using the convexity of \( f_t \), one can prove the desired result.
Online Newton Step

1. Choose $x_1 \in \mathbb{R}^d$, $\gamma, \epsilon > 0$ and $G_0 = \epsilon I \in \mathbb{R}^{d \times d}$.

2. In round $t$:
   2a. Compute $g_t = \nabla f_t(x_t)$,
   2b. Update $G_t = G_{t-1} + g_t g_t^\top$.
   2c. $x_{t+1} = \arg\min_{x \in \mathcal{X}} \|x - (x_t - \gamma G_t^{-1} g_t)\|_G^2$.

- Given $G_{t-1}$ and $g_t g_t^\top$, one can compute $G_t$ in time $O(d^2)$ using the following matrix inversion lemma [1] for invertible matrix $A$ and vector $x$,

$$
(A + xx^\top)^{-1} = A^{-1} - \frac{A^{-1}xx^\top A^{-1}}{1 + x^\top A^{-1}x}.
$$
Online Newton Step: Logarithmic Regret

**Exp-concavity property**

Let $\beta > 0$. A function $f$ is said to satisfy the $\beta$-exp-concavity property over $\mathcal{X}$ if

$$f(x) - f(y) \leq \nabla f(x)^\top (x - y) - \frac{\beta}{2} \left( \nabla f(x)^\top (x - y) \right)^2 \quad \forall x, y \in \mathcal{X}.$$ 

- The exp-concavity property is satisfied for a exp-concave function.
- It is also satisfied for a strongly convex function over a bounded domain.

**Theorem**

**Assumptions:**

- $f_1, \ldots, f_T$ satisfies the $\beta$-exp-concave property for some $\beta > 0$.
- $\max_{x, y \in \mathcal{X}} \|x - y\|_2 \leq b$.
- $\|g_t\|_2 \leq c$ for all $t$.

**Letting** $\gamma = \frac{1}{\beta}$, **then**

$$R_T \leq \frac{\beta \epsilon^2 b^2}{2} + \frac{\beta}{2} \log \left( c^2 T/\epsilon + 1 \right).$$

- Logarithmic regret bounds
Online Newton Step: Proof

- It is easy to prove that ONS can be rewritten as

\[
\begin{align*}
\nabla \psi_t(y_{t+1}) &= \nabla \psi_t(x_t) - g_t \\
 x_{t+1} &= \arg \min_{x \in X} B^{\psi_t}(x, y_{t+1}),
\end{align*}
\]

where \( \psi_t(\cdot) = \frac{1}{2\gamma} \| \cdot \|_{G_t}^2 \). Moreover, \( B^{\psi_t}(x, y) = \frac{1}{2\gamma} \| x - y \|_{G_t}^2 \), \( \| \cdot \|_t = \frac{1}{\sqrt{\gamma}} \cdot \| \cdot \|_{G_t} \)

and \( \| \cdot \|_{t,*} = \sqrt{\gamma} \cdot \| \cdot \|_{G_t^{-1}} \).

- Applying Proposition A, we get

\[
\sum_{t=1}^{T} \langle g_t, x_t - x \rangle \leq \frac{1}{2\gamma} \| x - x_1 \|_{G_1}^2 + \frac{1}{2\gamma} \sum_{t=2}^{T} \| x - x_t \|_{G_t}^2 - \| x - x_t \|_{G_{t-1}}^2 + \frac{\gamma}{2} \sum_{t=1}^{T} \| g_t \|_{G_{t-1}}^2.
\]

- Using the exp-concave property, and with \( \gamma = \beta^{-1} \),

\[
\sum_{t=1}^{T} f_t(x_t) - f_t(x) \leq \sum_{t=1}^{T} \langle g_t, x_t - x \rangle - \frac{\beta}{2} \left( \nabla f_t(x)^\top (x_t - x) \right)^2
\]

\[
\leq \frac{\beta}{2} \epsilon^2 \| x_1 - x \|_2^2 + \frac{\beta}{2} \sum_{t=1}^{T} \text{Tr} \left( G_{t-1}^{-1} g_t g_t^\top \right).
\]
• Now we only have to bound $\sum_{t \leq T} \text{Tr} \left( G_t^{-1} g_t g_t^\top \right)$:

$$
\sum_{t=1}^{T} \text{Tr} \left( G_t^{-1} g_t g_t^\top \right) = \sum_{t=1}^{T} \text{Tr} \left( G_t^{-1} (G_t - G_{t-1}) \right) \leq \sum_{t=1}^{T} \log \frac{|G_t|}{|G_{t-1}|} = \log \frac{|G_T|}{|G_0|}
$$

$$
\leq d \log ||G_T|| - \log |G_0| \leq d \log (c^2 T + \epsilon) - d \log \epsilon,
$$

where in the above we used the following lemma

**Lemma**

*Let $A \succeq B \succeq 0$. Then*

$$
\text{Tr} \left( A^{-1} (A - B) \right) \leq \log \frac{|A|}{|B|}
$$

**proof** Let $\lambda_i$ be the $i$-th eigenvalue of $A^{-1/2} B A^{-1/2}$.

$$
\text{Tr} \left( A^{-1} (A - B) \right) = \text{Tr} \left( A^{-1/2} (A - B) A^{-1/2} \right) = \text{Tr} \left( I - A^{-1/2} BA^{-1/2} \right)
$$

$$
= \sum_{i=1}^{d} (1 - \lambda_i) \leq \sum_{i=1}^{d} \log \lambda_i = \log \prod_{i=1}^{d} \lambda_i
$$

$$
= - \log |A^{-1/2} BA^{-1/2}| = \log \frac{|A|}{|B|}.
$$
Outline

Online Convex Optimization

Adaptive Online Mirror Descent (OMD)
  From OMD to Adaptive OMD
  Basic Analysis for Adaptive OMD
  Isotropically Adaptive Online Gradient Descent
  Diagonal AdaGrad
  Online Newton Step

Regret with Gradual Variations
Online Optimization with Gradual Variations

Adaptive Online Mirror Descent

1. Receive \( f_t \), compute \( g_t = \nabla f_t(x_t) \), and choose a mirror map \( \psi_t \),
2. \( x_{t+1} = \arg\min_{x \in \mathcal{X}} \{ \langle g_t, x \rangle + B^{\psi_t}(x, x_t) \} \).

What if we consider the following two-steps OMD?

META algorithm

1. Receive \( f_t \), compute \( g_t = \nabla f_t(x_t) \),
2. \( x_{t+1} = \arg\min_{x \in \mathcal{X}} \{ \langle g_t, x \rangle + B^{\psi_t}(x, x_t) \} \).
3. \( \tilde{x}_{t+1} = \arg\min_{x \in \mathcal{X}} \{ \langle \tilde{g}_t, x \rangle + B^{\psi_{t+1}}(x, x_{t+1}) \} \).

- "Mirror-prox" type online algorithms, with a subtle difference.
- \( \tilde{g}_t = 0 \rightarrow \) one-step OMD; \( \tilde{g}_t = \frac{1}{t} \sum_{s \leq t} g_s \) [5].
- If \( \tilde{g}_t \) is well chosen, (e.g. \( \tilde{g}_t \simeq \nabla f_t(\tilde{x}_{t+1}) \)), then the algorithm will perform better than standard OMD.
- In what follows we will study the case \( \tilde{g}_t = g_t \). [2]
Online Gradient Descent with Gradual Variations

\( \psi_t = \frac{1}{2\eta} \| \cdot \|_2^2 \)

<table>
<thead>
<tr>
<th><strong>META algorithm</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>- Receive ( f_t ), compute ( g_t = \nabla f_t(x_t) ),</td>
</tr>
<tr>
<td>- ( x_{t+1} = \arg \min_{x \in X} { \langle g_t, x \rangle + \frac{1}{2\eta} | x - x_t |_2^2 } ).</td>
</tr>
<tr>
<td>- ( \hat{x}<em>{t+1} = \arg \min</em>{x \in X} { \langle g_t, x \rangle + \frac{1}{2\eta} | x - x_{t+1} |_2^2 } ).</td>
</tr>
</tbody>
</table>

**Theorem**

Denote \( D_2 = \sum_{t \leq T} \max_{x \in X} \| \nabla f_t(x) - \nabla f_{t-1}(x) \|_2^2 \). Let \( \eta = 1/ \sqrt{D_2} \), and \( f_t \) be \( \lambda \)-smooth, with \( \lambda \leq 1/ \sqrt{8D_2} \). Then

\[ R_T \leq O(\sqrt{D_2}). \]
Online Newton Step with Gradual Variations

• \( \psi_t = \frac{1}{2}\| \cdot \|^2_{H_t} \), where \( H_t = (1 + \beta \gamma^2)I + \beta \sum_{s \leq t-1} g_t g_t^T \)

<table>
<thead>
<tr>
<th>META algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>. Receive ( f_t ), and ( g_t = \nabla f_t(x_t) ),</td>
</tr>
<tr>
<td>. ( x_{t+1} = \arg \min_{x \in X} { \langle g_t, x \rangle + | x - x_t |_{H_t}^2 } ).</td>
</tr>
<tr>
<td>. ( \hat{x}<em>{t+1} = \arg \min</em>{x \in X} { \langle g_t, x \rangle + | x - x_{t+1} |<em>{H</em>{t+1}}^2 } ).</td>
</tr>
</tbody>
</table>

• \( H_t \) is slightly different from the one in standard ONS.

**Theorem**

**Assumptions:**

- Denote \( D_2 = \sum_{t \leq T} \max_{x \in X} \| \nabla f_t(x) - \nabla f_{t-1}(x) \|^2_2 \) and let \( D_2 > 1 \).
- Each \( f_t \) satisfies \( \beta \)-exp-concave property and \( \lambda \)-smooth, with \( \beta \leq 1 \) and \( \lambda \geq 1 \).
- \( \max_{x \in X} \| x \|_2 \leq b \).

Then

\[
R_T \leq O(\beta b^2 + d\beta^{-1} \log(\lambda d D_2)).
\]
References

2005.

Online optimization with gradual variations.  

Adaptive subgradient methods for online learning and stochastic optimization.  

Logarithmic regret algorithms for online convex optimization.  

Extracting certainty from uncertainty: Regret bounded by variation in costs.  