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

Recommended reading materials

1. Hazan E, Agarwal A, Kale S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*. 2007, 69:169-92.
2. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*. 2011, 12: 2121-59.
3. Chiang CK, Yang T, Lee CJ, Mahdavi M, Lu CJ, Jin R, Zhu S. Online optimization with gradual variations. *Conference on Learning Theory*. 2012, 16: 6-1.
4. Gupta V, Koren T, Singer Y. A Unified Approach to Adaptive Regularization in Online and Stochastic Optimization. *arXiv:1706.06569*. 2017.

Online Convex Optimization

Online Convex Optimization

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

Example

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

Regret

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

$$R_T = \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^{T} f_t(\mathbf{x}).$$

Online Mirror Descent (OMD)

- Let $\psi : \mathcal{X} \to \mathbb{R}$ 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 ψ is given by

$$B^{\psi}(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Online Mirror Descent

1. Choose $\mathbf{x}_1 \in \mathbb{R}^d$, a mirror map ψ, and $\{\gamma_t > 0\}_{t=1}^{T}$.
2. In round t:
  2a. Receive f_t, compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$.
  2b. $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \{\langle \mathbf{g}_t, \mathbf{x} \rangle + \gamma_t^{-1} B^{\psi}(\mathbf{x}, \mathbf{x}_t)\}$.

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

Convergence

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

$$R_T \leq \left(\frac{1}{\gamma} \max_{\mathbf{x} \in \mathcal{X}, t \leq T} B^{\psi}(\mathbf{x}, \mathbf{x}_t) + \frac{\gamma}{2} \max_{t \leq T} \|\mathbf{g}_t\|_*^2 \right) \sqrt{T}$$

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

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 $\mathbf{x}_1 \in \mathbb{R}^d$, a mirror map ψ and $\{\gamma_t > 0\}_{t=1}^T$.
2. In round t:
2a. Receive f_t, compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$.
2b. $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{g}_t, \mathbf{x}\rangle + \gamma_t^{-1} B^{\psi}(\mathbf{x}, \mathbf{x}_t)\}$.

- Fixed mirror map ψ, (possibly) time-varying step-size γ_t.

Generalized Online Mirror Descent

1. Choose $\mathbf{x}_1 \in \mathbb{R}^d$.
2. In round t:
2a. Receive f_t, compute $g_t = \nabla f_t(\mathbf{x}_t)$, and choose a mirror map ψ_t,
2b. $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\mathbf{g}_t, \mathbf{x}\rangle + B^{\psi_t}(\mathbf{x}, \mathbf{x}_t)\}$.

- ψ_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

Adaptive Online Mirror Descent

1. Choose $\mathbf{x}_1 \in \mathbb{R}^d$.
2. In round t:
  2a. Receive f_t, compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$, and choose a mirror map ψ_t,
  2b. $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{g}_t, \mathbf{x}\rangle + B^{\psi_t}(\mathbf{x}, \mathbf{x}_t)\}$.

- ψ_t is an (adaptive) regularization function which may depend on $\{(\mathbf{x}_s, \mathbf{g}_s)\}_{s\le t}$.

Adaptive Strategies for OMD

In what follows, we will study the following three different choices of ψ_t.

Isotropically Adaptive OGD

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

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

Diagonal AdaGrad [3]

Let $\mathbf{G}_t = \epsilon \boldsymbol{I} + \sum_{s\leq t} \mathbf{g}_s\mathbf{g}_s^\top$ and $\mathbf{S}_t = \mathrm{diag}(\mathbf{G}_t)$. Let $\max_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x}-\mathbf{y}\|_\infty \leq b_\infty$.

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

Online Newton Step [4]

Let f_t be β-exp-concave. Choose $\psi_t = \beta\|\cdot\|_{\mathbf{G}_t}^2/2$. Then $R_T \lesssim \log T$.

*Some Basic Notations and Lemmas

Strong Convexity

A function $\psi : \mathcal{X} \rightarrow \mathbb{R}$ is μ-strongly convex w.r.t. a norm $\|\cdot\|$ if for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

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

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 $\mathbf{H} \in \mathbb{R}^{d\times d}$, $\psi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{H}\mathbf{x}$ is 1-strongly convex w.r.t. the norm $\|\cdot\|_{\mathbf{H}} = \sqrt{\mathbf{x}^\top\mathbf{H}\mathbf{x}}$.

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

$$\langle \nabla f(\hat{\mathbf{x}}), \mathbf{y} - \hat{\mathbf{x}}\rangle \geq 0, \quad \forall \mathbf{y} \in \mathcal{X}.$$

*Bregman divergences

Definition (Bregman divergence)

Let $\psi : \mathcal{X} \to \mathbb{R}$ 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^ψ) associated with ψ for points $\mathbf{x}$ and $\mathbf{y}$ is:

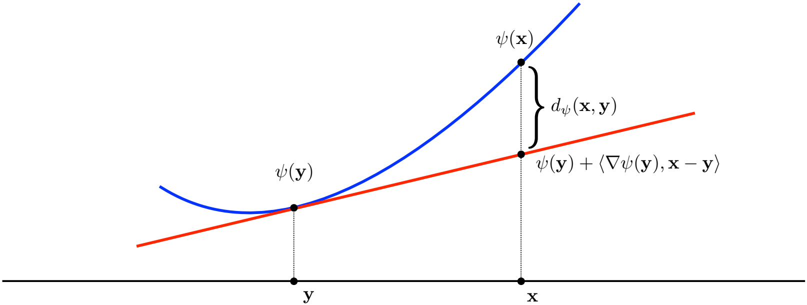
$$B^\psi(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

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

- $B^\psi(\mathbf{x}, \mathbf{y}) \neq B^\psi(\mathbf{y}, \mathbf{x})$ in general.

*Bregman divergences

- The Bregman divergence is the vertical distance at $\mathbf{x}$ between ψ and the tangent of ψ at $\mathbf{y}$, 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 $B^\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{y}\|_2^2$.
- Let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be positive definite. $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$. Then $\psi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{A}\mathbf{x}$ is strongly-convex and $B^\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_{\mathbf{A}}^2$.

Equivalent Form for Mirror Descent Step

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

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{g}_t, \mathbf{x}\rangle + B^{\psi_t}(\mathbf{x}, \mathbf{x}_t)\}. \tag{1}$$

Lemma (A)

Equation (1) is equivalent to[1]

$$\begin{cases} \nabla\psi_t(\mathbf{y}_{t+1}) = \nabla\psi_t(\mathbf{x}_t) - \mathbf{g}_t \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} B^{\psi_t}(\mathbf{x}, \mathbf{y}_{t+1}). \end{cases}$$

Proof. The proof is straightforward.

[1] Here, we assume the existence of $\mathbf{y}_{t+1}$ such that it satisfies the first equation. The existence of $\mathbf{y}_{t+1}$ is always guaranteed if there exists a domain $\mathcal{D}$ such that $\mathcal{X} \subseteq \mathcal{D}$ and $\nabla\psi_t(\mathcal{D}) = \mathbb{R}^d$.

Basic Property for Mirror Descent Step

$$\begin{cases} \psi_t(\mathbf{y}_{t+1}) = \nabla\psi_t(\mathbf{x}_t) - \mathbf{g}_t \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} B^{\psi_t}(\mathbf{x}, \mathbf{y}_{t+1}). \end{cases} \tag{1}$$

Lemma A

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

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x}\rangle \leq B^{\psi_t}(\mathbf{x}, \mathbf{x}_t) - B^{\psi_t}(\mathbf{x}, \mathbf{x}_{t+1}) - B^{\psi_t}(\mathbf{x}_{t+1}, \mathbf{x}_t).$$

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

$$\begin{aligned} &\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x}\rangle = \langle \nabla\psi_t(\mathbf{x}_t) - \nabla\psi_t(\mathbf{y}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x}\rangle \\ &= \langle \nabla\psi_t(\mathbf{x}_t) - \nabla\psi_t(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x}\rangle + \langle \nabla\psi_t(\mathbf{x}_{t+1}) - \nabla\psi_t(\mathbf{y}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x}\rangle. \end{aligned}$$

Since $\mathbf{x}_{t+1}$ is the minimizer, according to the first order optimality condition,

$$0 \geq \langle \nabla B^{\psi_t}(\mathbf{x}, \mathbf{y}_{t+1})|_{\mathbf{x}=\mathbf{x}_{t+1}}, \mathbf{x}_{t+1} - \mathbf{x}\rangle = \langle \nabla\psi_t(\mathbf{x}_{t+1}) - \nabla\psi_t(\mathbf{y}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x}\rangle.$$

Thus,

$$\begin{aligned} \langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x}\rangle &\leq \langle \nabla\psi_t(\mathbf{x}_t) - \nabla\psi_t(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x}\rangle \\ &= B^{\psi_t}(\mathbf{x}, \mathbf{x}_t) - B^{\psi_t}(\mathbf{x}, \mathbf{x}_{t+1}) - B^{\psi_t}(\mathbf{x}_{t+1}, \mathbf{x}_t). \end{aligned}$$

$$\begin{cases} \nabla\psi_t(\mathbf{y}_{t+1}) = \nabla\psi_t(\mathbf{x}_t) - \mathbf{g}_t \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} B^{\psi_t}(\mathbf{x}, \mathbf{y}_{t+1}). \end{cases}$$

Lemma A

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

$$\langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x} \rangle \leq B^{\psi_t}(\mathbf{x}, \mathbf{x}_t) - B^{\psi_t}(\mathbf{x}, \mathbf{x}_{t+1}) - B^{\psi_t}(\mathbf{x}_{t+1}, \mathbf{x}_t).$$

By Cauchy-Swartz inequality,

$$\begin{aligned} &\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle = \langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x} \rangle + \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \\ &\leq \langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x} \rangle + \|\mathbf{g}_t\|_{t,*} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_t \\ &\leq \langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{g}_t\|_{t,*}^2 + \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_t^2 \\ &\leq \langle \mathbf{g}_t, \mathbf{x}_{t+1} - \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{g}_t\|_{t,*}^2 + \frac{1}{2} B^{\psi_t}(\mathbf{x}_{t+1}, \mathbf{x}_t), \end{aligned}$$

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

$$\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle \leq B^{\psi_t}(\mathbf{x}, \mathbf{x}_t) - B^{\psi_t}(\mathbf{x}, \mathbf{x}_{t+1}) + \frac{1}{2}\|\mathbf{g}_t\|_{t,*}^2,$$

So far, we have proved that

$$\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle \leq B^{\psi_t}(\mathbf{x}, \mathbf{x}_t) - B^{\psi_t}(\mathbf{x}, \mathbf{x}_{t+1}) + \frac{1}{2}\|\mathbf{g}_t\|_{t,*}^2,$$

Writing $B^{\psi_t}(\mathbf{x}, \mathbf{x}_t) - B^{\psi_t}(\mathbf{x}, \mathbf{x}_{t+1})$ as

$$\Big(B^{\psi_t}(\mathbf{x}, \mathbf{x}_t) - B^{\psi_{t+1}}(\mathbf{x}, \mathbf{x}_{t+1})\Big) + \Big(B^{\psi_{t+1}}(\mathbf{x}, \mathbf{x}_{t+1}) - B^{\psi_t}(\mathbf{x}, \mathbf{x}_{t+1})\Big)$$

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

$$\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle \leq B^{\psi_1}(\mathbf{x}, \mathbf{x}_1) - B^{\psi_T}(\mathbf{x}, \mathbf{x}_{T+1}) + \sum_{t=1}^{T-1} \Big(B^{\psi_{t+1}}(\mathbf{x}, \mathbf{x}_{t+1}) - B^{\psi_t}(\mathbf{x}, \mathbf{x}_{t+1})\Big) + \frac{1}{2}\|\mathbf{g}_t\|_{t,*}^2,$$

Noting that $B^{\psi_T}(\mathbf{x}, \mathbf{x}_{T+1}) \geq 0$, we thus prove the following result.

Basic Property of Mirror Descent

Proposition A

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

$$\begin{cases} \nabla\psi_t(\mathbf{y}_{t+1}) = \nabla\psi_t(\mathbf{x}_t) - \mathbf{g}_t \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} B^{\psi_t}(\mathbf{x}, \mathbf{y}_{t+1}). \end{cases}$$

Then

$$\sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x} \rangle \leq B^{\psi_1}(\mathbf{x}, \mathbf{x}_1) + \sum_{t=2}^{T} B^{\psi_t}(\mathbf{x}, \mathbf{x}_t) - B^{\psi_{t-1}}(\mathbf{x}, \mathbf{x}_t) + \frac{1}{2}\sum_{t=1}^{T} \|\mathbf{g}_t\|_{t,*}^2$$

Isotropically Adaptive Online Gradient Descent (Isotropic Adagrad)

Adaptive Online Gradient Descent

1. Choose $\mathbf{x}_1 \in \mathbb{R}^d$ and $\gamma > 0$.
2. In round t:

2a. Compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$ and $\gamma_t = \gamma / \sqrt{\sum_{s \le t} \|\mathbf{g}_s\|_2^2}$,

2b. $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - (\mathbf{x}_t - \gamma_t \mathbf{g}_t)\|_2$.

- w.l.o.g. we assume that $\|\mathbf{g}_1\|_2 > 0$.

Convergence

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

$$R_T \le (\gamma^{-1} b^2/2 + \gamma) \sqrt{\sum_{s=1}^{T} \|\mathbf{g}_s\|_2^2}$$

- Adaptive learning rate.

Proof.

Let $\psi_t = \frac{1}{2\gamma_t}\|\cdot\|_2^2$. It is easy to see that ψ is strongly convex with respect to $\|\cdot\|_t = \frac{1}{\sqrt{\gamma_t}}\|\cdot\|_2$. Moreover, $B^{\psi_t}(\mathbf{x},\mathbf{y}) = \frac{1}{2\gamma_t}\|\mathbf{x}-\mathbf{y}\|_2^2$, 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

$$\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}\rangle \leq \frac{1}{2\gamma_1}\|\mathbf{x}-\mathbf{x}_1\|^2 + \sum_{t=2}^{T}\|\mathbf{x}-\mathbf{x}_t\|_2^2\left(\frac{1}{2\gamma_t} - \frac{1}{2\gamma_{t-1}}\right) + \frac{1}{2}\sum_{t=1}^{T}\gamma_t\|\mathbf{g}_t\|_2^2$$

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

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

$$f_t(\mathbf{x}_t) - f(\mathbf{x}) \leq \langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}\rangle,$$

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

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - (\mathbf{x}_t - \gamma_t \mathbf{g}_t)\|_2^2.$$

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

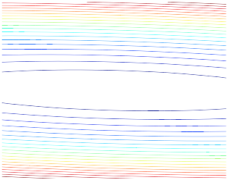
$$\mathbf{x}_{t+1}(i) = \mathbf{x}_t(i) - \gamma_t \mathbf{g}_t(i).$$

All feature dimension share the same learning rate.

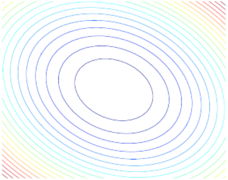
- Practical examples often have high-dimensional feature spaces.
– Many features are irrelevant
– Rare features are often very informative.

Diagonal AdaGrad: Motivation

Why adapt to geometry?



Hard



Nice

y_t	$\phi_{t,1}$	$\phi_{t,2}$	$\phi_{t,3}$
1	1	0	0
-1	.5	0	1
1	-.5	1	0
-1	0	0	0
1	.5	0	0
-1	1	0	0
1	-1	1	0
-1	-.5	0	1

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

Examples from Duchi et al. ISMP 2012 slides

Diagonal AdaGrad: Motivation

- Standard online gradient descent:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - (\mathbf{x}_t - \gamma \mathbf{g}_t)\|_2^2.$$

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

$$R_T \leq \frac{1}{2\gamma}\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}_1\|_2^2 + \frac{\gamma}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_2^2$$

- What if we consider $\psi_t = \frac{1}{2\gamma}\|\cdot\|_{\mathbf{H}}^2$?

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - (\mathbf{x}_t - \gamma \mathbf{H}^{-1}\mathbf{g}_t)\|_{\mathbf{H}}^2$$

The regret bound becomes:

$$R_T \leq \frac{1}{2\gamma}\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}_1\|_2^2 + \frac{\gamma}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\mathbf{H}^{-1}}^2.$$

What $\mathbf{H}$ to choose?

Diagonal AdaGrad: Motivation

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - (\mathbf{x}_t - \gamma \mathbf{H}^{-1}\mathbf{g}_t)\|_{\mathbf{H}}^2$$

- The regret bound:

$$R_T \leq \frac{1}{2\gamma} \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}_1\|_2^2 + \frac{\gamma}{2} \sum_{t=1}^{T} \|\mathbf{g}_t\|_{\mathbf{H}^{-1}}^2 .$$

- Choosing $\mathbf{H}$ to minimize

$$\sum_{t=1}^{T} \|\mathbf{g}_t\|_{\mathbf{H}^{-1}}^2 = \mathrm{Tr}\left(\mathbf{H}^{-1} \sum_{t=1}^{T} \mathbf{g}_t \mathbf{g}_t^\top\right),$$

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

$$\mathbf{H} = c \left(\sum_{t\leq T} \mathbf{g}_t \mathbf{g}_t^\top\right)^{1/2}$$

Diagonal AdaGrad: Motivation

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - (\mathbf{x}_t - \gamma \mathbf{H}^{-1}\mathbf{g}_t)\|_{\mathbf{H}}^2$$

- The regret bound:

$$R_T \leq \frac{1}{2\gamma} \max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}_1\|_2^2 + \frac{\gamma}{2} \sum_{t=1}^{T} \|\mathbf{g}_t\|_{\mathbf{H}^{-1}}^2.$$

- Choosing $\mathbf{H}$ to minimize

$$\sum_{t=1}^{T} \|\mathbf{g}_t\|_{\mathbf{H}^{-1}}^2 = \mathrm{Tr}\left(\mathbf{H}^{-1} \sum_{t=1}^{T} \mathbf{g}_t \mathbf{g}_t^\top\right),$$

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

$$\mathbf{H} = c\left(\sum_{t\leq T} \mathbf{g}_t \mathbf{g}_t^\top\right)^{1/2}$$

- At time t, replaced $\mathbf{H}$ by

$$\mathbf{H}_t = \left(\sum_{s\leq t} \mathbf{g}_s \mathbf{g}_s^\top\right)^{1/2}$$

and to reduce the computational cost, replace $\mathbf{H}_t$ by $\mathrm{diag}(\mathbf{H}_t)$

Diagonal AdaGrad

AdaGrad

1. Choose $\mathbf{x}_1 \in \mathbb{R}^d$ and $\gamma > 0$.
2. In round t:
  2a. Compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$,
  2b. Update $\mathbf{A}_t = [\mathbf{A}_{t-1}\ \mathbf{g}_t], \boldsymbol{s}_t(i) = \|\mathbf{A}_t(i,:)\|_2^2$, $\mathbf{S}_t = (\epsilon \boldsymbol{I} + \mathrm{diag}(\boldsymbol{s}_t))^{1/2}$.
  2c. $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - (\mathbf{x}_t - \gamma \mathbf{S}_t^{-1}\mathbf{g}_t)\|_{\mathbf{S}_t}^2$.

Simple cases

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

$$\mathbf{x}_{t+1}(i) = \mathbf{x}_t(i) - \frac{\gamma}{\sqrt{\epsilon + \sum_{s\leq t} |\mathbf{g}_s(i)|^2}} \mathbf{g}_t(i)$$

• Each feature dimension has it's own learning rate. The learning rate is adapted with t and takes geometry of the past observations into account.
• When $|\mathbf{g}_s(i)|^2 \approx 0$ for most s, then $\sum_{s\leq t} |\mathbf{g}_t(i)|^2 \approx 0$. The learning rate in i-th feature tends to be larger, meaning that the i-the feature is important.

Diagonal AdaGrad: Convergence

AdaGrad

1. Choose $\mathbf{x}_1 \in \mathbb{R}^d$ and $\gamma > 0$.
2. In round t:
2a. Compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$,
2b. Update $\mathbf{A}_t = [\mathbf{A}_{t-1}\ \mathbf{g}_t], \boldsymbol{s}_t(i) = \|\mathbf{A}_t(i,:)\|_2^2$, $\mathbf{S}_t = (\epsilon \boldsymbol{I} + \mathrm{diag}(\boldsymbol{s}_t))^{1/2}$.
2c. $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - (\mathbf{x}_t - \gamma \mathbf{S}_t^{-1}\mathbf{g}_t)\|_{\mathbf{S}_t}^2$.

Convergence

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

$$R_T \le (b_\infty^2/(2\gamma) + \gamma)\mathrm{Tr}\,(\mathbf{S}_T)$$

- Letting $\epsilon \to 0$, $\mathrm{Tr}\,(\mathbf{S}_T) \to \mathrm{Tr}\,(\mathrm{diag}(\boldsymbol{s}_t)) = \sum_{i=1}^d \sqrt{\sum_{t=1}^T |\mathbf{g}_t(i)|^2}$

AdaGrad Theoretical Examples

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

- SVM example. Let $f_t(x) = (1 - b_t \mathbf{a}_t^\top \mathbf{x})$ with $\mathbf{a}_t \in \{-1, 1, 0\}^d$. If $\mathbf{a}_t(i) \neq 0$ with probability $\propto i^{-\alpha}$ where $\alpha > 1$. Then

$$\mathbb{E} R_T(\mathbf{x}) \lesssim \sqrt{T} \max(\log d, d^{1-\alpha/2}).$$

- Previously regret bound:

$$\mathbb{E} R_T(\mathbf{x}) \lesssim \sqrt{Td}.$$

Diagonal AdaGrad: Theoretical Analysis

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

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

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

$$\begin{cases} \nabla\psi_t(\mathbf{y}_{t+1}) = \nabla\psi_t(\mathbf{x}_t) - \mathbf{g}_t \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} B^{\psi_t}(\mathbf{x},\mathbf{y}_{t+1}). \end{cases}$$

- Applying Proposition A,

$$\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{x}_t - \mathbf{x}\rangle \le \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{x}_1\|^2_{\mathbf{S}_1} + \frac{1}{2\gamma}\sum_{t=2}^{T}\|\mathbf{x}-\mathbf{x}_t\|^2_{\mathbf{S}_t} - \|\mathbf{x}-\mathbf{x}_t\|^2_{\mathbf{S}_{t-1}} + \frac{\gamma}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2_{\mathbf{S}_t^{-1}}$$

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

-

$$\frac{1}{2\gamma}\|\mathbf{x}-\mathbf{x}_1\|^2_{\mathbf{S}_1} = \frac{1}{2\gamma}(\mathbf{x}-\mathbf{x}_1)^\top \mathbf{S}_1(\mathbf{x}-\mathbf{x}_1) \le \frac{\|\mathbf{x}-\mathbf{x}_1\|^2_\infty}{2\gamma}\mathrm{Tr}\left(\mathbf{S}_1\right) \le \frac{b^2_\infty}{2\gamma}\mathrm{Tr}\left(\mathbf{S}_1\right)$$

- $$\frac{1}{2\gamma}\sum_{t=2}^{T}\|\mathbf{x}-\mathbf{x}_t\|_{\mathbf{S}_t}^2-\|\mathbf{x}-\mathbf{x}_t\|_{\mathbf{S}_{t-1}}^2=\frac{1}{2\gamma}\sum_{t=2}^{T}(\mathbf{x}-\mathbf{x}_t)^\top(\mathbf{S}_t-\mathbf{S}_{t-1})(\mathbf{x}-\mathbf{x}_t)$$
$$\leq\frac{1}{2\gamma}\sum_{t=2}^{T}\|\mathbf{x}-\mathbf{x}_t\|_\infty^2\,\mathrm{Tr}\left(\mathbf{S}_t-\mathbf{S}_{t-1}\right)\leq\frac{b_\infty^2}{2\gamma}\mathrm{Tr}\left(\mathbf{S}_T-\mathbf{S}_1\right).$$

- $$\sum_{t=1}^{T}\|\mathbf{g}_t\|_{\mathbf{S}_t^{-1}}^2=\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{|\mathbf{g}_t(i)|^2}{\boldsymbol{s}_t(i)}=\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{|\mathbf{g}_t(i)|^2}{\sqrt{\epsilon\boldsymbol{I}+\sum_{j=1}^{t}|\mathbf{g}_j(i)|}}$$
$$\leq 2\sum_{i=1}^{d}\sqrt{\sum_{t=1}^{T}|\mathbf{g}_t(i)|^2},$$

where for the last inequality we used the basic inequality $2\sum_{t\leq T}a_t(\sum_{s\leq t}a_s)^{-1/2}\leq(\sum_{t\leq T}a_t)^{1/2}$.
Combining the above estimates and using the convexity of f_t, one can prove the desired result.

Online Newton Step

Online Newton Step

1. Choose $\mathbf{x}_1 \in \mathbb{R}^d$, $\gamma, \epsilon > 0$ and $\mathbf{G}_0 = \epsilon \boldsymbol{I} \in \mathbb{R}^{d \times d}$.
2. In round t:
   - **2a.** Compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$,
   - **2b.** Update $\mathbf{G}_t = \mathbf{G}_{t-1} + \mathbf{g}_t \mathbf{g}_t^\top$.
   - **2c.** $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - (\mathbf{x}_t - \gamma \mathbf{G}_t^{-1} \mathbf{g}_t)\|_{\mathbf{G}_t}^2$.

• Given $\mathbf{G}_{t-1}$ and $\mathbf{g}_t \mathbf{g}_t^\top$, one can compute $\mathbf{G}_t$ in time $O(d^2)$ using the following matrix inversion lemma [1] for invertible matrix $\mathbf{A}$ and vector $\mathbf{x}$,

$$(\mathbf{A} + \mathbf{x}\mathbf{x}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{x}^\top\mathbf{A}^{-1}}{1 + \mathbf{x}^\top\mathbf{A}^{-1}\mathbf{x}}.$$

Online Newton Step: Logarithmic Regret

Exp-concavity property

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

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

- 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:

- $f_1, \cdots, f_T$ *satisfies the* β*-exp-concave property for some* $\beta > 0$.
- $\max_{\mathbf{x}, \mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|_2 \leq b$.
- $\|\mathbf{g}_t\|_2 \leq c$ *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

$$\begin{cases} \nabla\psi_t(\mathbf{y}_{t+1}) = \nabla\psi_t(\mathbf{x}_t) - \mathbf{g}_t \\ \mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} B^{\psi_t}(\mathbf{x},\mathbf{y}_{t+1}), \end{cases}$$

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

- Applying Proposition A, we get

$$\sum_{t=1}^{T}\langle\mathbf{g}_t,\mathbf{x}_t-\mathbf{x}\rangle \le \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{x}_1\|^2_{\mathbf{G}_1} + \frac{1}{2\gamma}\sum_{t=2}^{T}\|\mathbf{x}-\mathbf{x}_t\|^2_{\mathbf{G}_t} - \|\mathbf{x}-\mathbf{x}_t\|^2_{\mathbf{G}_{t-1}} + \frac{\gamma}{2}\sum_{t=1}^{T}\|\mathbf{g}_t\|^2_{\mathbf{G}_t^{-1}}.$$

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

$$\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}) \le \sum_{t=1}^{T}\langle\mathbf{g}_t,\mathbf{x}_t-\mathbf{x}\rangle - \frac{\beta}{2}\left(\nabla f_t(\mathbf{x})^\top(\mathbf{x}_t-\mathbf{x})\right)^2$$

$$\le \frac{\beta}{2}\epsilon^2\|\mathbf{x}_1-\mathbf{x}\|_2^2 + \frac{\beta}{2}\sum_{t=1}^{T}\mathrm{Tr}\left(\mathbf{G}_t^{-1}\mathbf{g}_t\mathbf{g}_t^\top\right).$$

- Now we only have to bound $\sum_{t \le T} \operatorname{Tr}\left(\mathbf{G}_t^{-1}\mathbf{g}_t\mathbf{g}_t^\top\right)$:

$$\sum_{t=1}^{T} \operatorname{Tr}\left(\mathbf{G}_t^{-1}\mathbf{g}_t\mathbf{g}_t^\top\right) = \sum_{t=1}^{T} \operatorname{Tr}\left(\mathbf{G}_t^{-1}(\mathbf{G}_t - \mathbf{G}_{t-1})\right) \le \sum_{t=1}^{T} \log \frac{|\mathbf{G}_t|}{|\mathbf{G}_{t-1}|} = \log \frac{|\mathbf{G}_T|}{|\mathbf{G}_0|}$$
$$\le d \log \|\mathbf{G}_T\| - \log |\mathbf{G}_0| \le d \log(c^2 T + \epsilon) - d \log \epsilon,$$

where in the above we used the following lemma

Lemma

Let $\mathbf{A} \succeq \mathbf{B} \succeq 0$. *Then*

$$\operatorname{Tr}\left(\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\right) \le \log \frac{|\mathbf{A}|}{|\mathbf{B}|}$$

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

$$\operatorname{Tr}\left(\mathbf{A}^{-1}(\mathbf{A} - \mathbf{B})\right) = \operatorname{Tr}\left(\mathbf{A}^{-1/2}(\mathbf{A} - \mathbf{B})\mathbf{A}^{-1/2}\right) = \operatorname{Tr}\left(\boldsymbol{I} - \mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}\right)$$
$$= \sum_{i=1}^{d}(1 - \lambda_i) \le -\sum_{i=1}^{d} \log \lambda_i = \log \prod_{i=1}^{d} \lambda_i$$
$$= -\log |\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}| = \log \frac{|\mathbf{A}|}{|\mathbf{B}|}.$$

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

- Receive f_t, compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$, and choose a mirror map ψ_t,
- $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{g}_t, \mathbf{x}\rangle + B^{\psi_t}(\mathbf{x}, \mathbf{x}_t)\}$.

What if we consider the following two-steps OMD?

META algorithm

- Receive f_t, compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$,
- $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{g}_t, \mathbf{x}\rangle + B^{\psi_t}(\mathbf{x}, \mathbf{x}_t)\}$.
- $\hat{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle \tilde{\mathbf{g}}_t, \mathbf{x}\rangle + B^{\psi_{t+1}}(\mathbf{x}, \mathbf{x}_{t+1})\}$.

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

Online Gradient Descent with Gradual Variations

- $\psi_t = \frac{1}{2\eta}\|\cdot\|_2^2$

META algorithm

- Receive f_t, compute $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$,
- $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{g}_t, \mathbf{x}\rangle + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_t\|_2^2\}$.
- $\hat{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle \mathbf{g}_t, \mathbf{x}\rangle + \frac{1}{2\eta}\|\mathbf{x} - \mathbf{x}_{t+1}\|_2^2\}$.

Theorem

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

$$R_T \le O(\sqrt{D_2}).$$

Online Newton Step with Gradual Variations

- $\psi_t = \frac{1}{2}\|\cdot\|^2_{\mathbf{H}_t}$, where $\mathbf{H}_t = (1+\beta\gamma^2)\boldsymbol{I} + \beta\sum_{s\leq t-1}\mathbf{g}_t\mathbf{g}_t^\top$

META algorithm

- Receive f_t, and $\mathbf{g}_t = \nabla f_t(\mathbf{x}_t)$,
- $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle\mathbf{g}_t, \mathbf{x}\rangle + \|\mathbf{x}-\mathbf{x}_t\|^2_{\mathbf{H}_t}\}$.
- $\hat{\mathbf{x}}_{t+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{\langle\mathbf{g}_t, \mathbf{x}\rangle + \|\mathbf{x}-\mathbf{x}_{t+1}\|^2_{\mathbf{H}_{t+1}}\}$.

- $\mathbf{H}_t$ is slightly different from the one in standard ONS.

Theorem

Assumptions:

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

Then

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

References I

[1] Mike Brookes.
The Matrix Reference Manual.
2005.

[2] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu.
Online optimization with gradual variations.
In *Conference on Learning Theory*, pages 6–1, 2012.

[3] John Duchi, Elad Hazan, and Yoram Singer.
Adaptive subgradient methods for online learning and stochastic optimization.
Journal of Machine Learning Research, 12(Jul):2121–2159, 2011.

[4] Elad Hazan, Amit Agarwal, and Satyen Kale.
Logarithmic regret algorithms for online convex optimization.
Machine Learning, 69(2):169–192, 2007.

[5] Elad Hazan and Satyen Kale.
Extracting certainty from uncertainty: Regret bounded by variation in costs.
Machine learning, 80(2):165–188, 2010.