Adaptive Subgradient Methods for Online Learning and Stochastic Optimization

Journal of Machine Learning Research (2011)

John Duchi, Elad Hazan and Yoram Singer

Presenter: Gyuseung Baek

June 3, 2019

Introduction

- New family of subgradient methods that dynamically incorporate knowledge of the geometry of the data
- Boost learning rarely observed variable's coefficient.

Existing online learning algorithm

- Suppose h is 1-strongly convex function.
- RDA: update $\{x_t\}$ as

$$x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \{\eta \langle \bar{g}_t, x \rangle + \eta\psi(x) + h_t(x)\} \tag{1}$$

- FOBOS (Forward-backward spliiting): update $\{x_t\}$ as
  - $x_{t+\frac{1}{2}} = x_t - \alpha_t g_t$
  - $\arg\min_x \left\{ \frac{1}{2} \left\| x - x_{t+\frac{1}{2}} \right\|_2^2 + \alpha_t \psi(x) \right\}$
- Generalization of FOBOS : update $\{x_t\}$ as

$$x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \{\eta \langle g_t, x \rangle + \eta\psi(x) + B_{h_t}(x, x_t)\} \tag{2}$$

where $B_{h_t}(\cdot)$: Bregman divergence associated with h_t. Let $||\cdot||_t$ be a norm which make h_t 1-strongly convex.

Regret bound of existing online learning algorithm

- **Proposition 2** Let the sequence $\{x_t\}$ be defined by (1). For any $x \in \mathcal{X}$,

$$R_t(x) \leq \frac{1}{\eta} h_t(x) + \frac{\eta}{2} \sum_{\tau=1}^{t} \|g_\tau\|_{\tau-1}^2 \tag{3}$$

- **Proposition 3** Let the sequence $\{x_t\}$ be defined by (2). Assume w.l.o.g. that $\psi(x_1) = 0$. For any $x \in \mathcal{X}$,

$$\begin{aligned} R_t(x) &\leq \frac{1}{\eta} B_{h_1}(x, x_1) \\ &+ \frac{1}{\eta} \sum_{\tau=1}^{t-1} \left[B_{h_{\tau+1}}(x, x_{\tau+1}) - B_{h_\tau}(x, x_{\tau+1}) \right] + \frac{\eta}{2} \sum_{\tau=1}^{t} \|g_t\|_{h_\tau}^2 \end{aligned} \tag{4}$$

Intuition of AdaGrad

- In regret bounds, selection of h_t effect to the norm $\|\cdot\|_{h_t}$.
- If h_t satisfies $\|x\|_{h_t} = \langle x, diag(s)^{-1}x\rangle$ for some s (it holds if $h_t(x) = \langle x, diag(s)x\rangle$), then

$$\sum_{\tau=1}^{t} \|g_\tau\|_{\tau-1}^2 = \sum_{\tau=1}^{t}\sum_{i=1}^{d} \frac{g_{t,i}^2}{s_i} \tag{5}$$

- For $s \succeq 0, \langle 1, s\rangle \leq c$, equation (5) is minimized if $s = c' \cdot g_{1:t}$ where c' is a normalized constant.

Regret bound of AdaGrad

- **Thm 5** Let $h_t(x) = \langle x, diag(\delta + g_{1:t})x \rangle$ and suppose $\delta \geq \max_t \|g_t\|_\infty$. Then for any $x \in \mathcal{X}$, $\{x_t\}$ acquired by (1) has regret bound

$$\begin{aligned} R_t(x) &\leq \frac{\delta}{\eta}\|x\|_2^2 + \frac{1}{\eta}\|x\|_\infty^2 \sum_{i=1}^{d} \|g_{1:t,i}\|_2 + \eta \sum_{i=1}^{d} \|g_{1:t,i}\|_2 \\ &= O\left(\|x\|_\infty \sum_{i=1}^{d} \|g_{1:t,i}\|_2\right) \end{aligned} \tag{6}$$

  For any $x \in \mathcal{X}$, $\{x_t\}$ acquired by (2) has regret bound

$$\begin{aligned} R_t(x) &\leq \frac{1}{2\eta}\max_{\tau \leq t}\|x - x_t\|_\infty^2 \sum_{i=1}^{d} \|g_{1:t,i}\|_2 + \eta \sum_{i=1}^{d} \|g_{1:t,i}\|_2 \\ &= O\left(\max_{\tau \leq t}\|x_\tau - x\|_\infty \sum_{i=1}^{d} \|g_{1:t,i}\|_2\right) \end{aligned} \tag{7}$$

Strength of AdaGrad

- If domain $\mathcal{X}$ is bouned by infinity norm, 2-norm is much bigger than ∞-norm. So it can avoid the curse of dimensionality
- If variable i is not observed before time t, $g_{\tau,i} = 0$ for all $\tau \leq t$. So coefficients of variable i is rapidly learned at time t. (This is proven by experiments)

Proof of Proposition 2

Define h_t^* to be the conjugate dual of $\psi(x) + h_t(x)/\eta$

$$h_t^*(g) = \sup_{x \in \mathcal{X}} \left\{ \langle g, x \rangle - \psi(x) - \frac{1}{\eta} h_t(x) \right\}$$

Since ψ_t/η is $1/\eta$strongly convex with respect to the norm $\|\cdot\|_{h_t}$, the function h_t^* has η-Lipschitz continuous gradients with respect to $\|\cdot\|_{h_t}$

$$\|\nabla h_t^*(g_1) - \nabla h_t^*(g_2)\|_{\psi_t} \leq \eta \|g_1 - g_2\|_{h_t^*} \tag{8}$$

for any g_1, g_2.
Futher, a simple argument with the fundatmental theorem of calculus gives that if g has L-Lipschitz gradients,
$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + (L/2)\|y - x\|^2$, and

$$\nabla h_t^*(g) = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ -\langle g, x \rangle + \psi(x) + \frac{1}{\eta} h_t(x) \right\} \tag{9}$$

Proof of Proposition 2

Then

$$
\begin{aligned}
R_t(x) &= \sum_{\tau=1}^{t} f_\tau\left(x_\tau\right) + \psi\left(x_\tau\right) - f_\tau\left(x\right) - \psi\left(x\right) \\
&\leq \sum_{\tau=1}^{t} \left\langle g_\tau, x_\tau - x \right\rangle - \psi\left(x\right) + \psi\left(x_t\right) \\
&\leq \sum_{\tau=1}^{t} \left\langle g_\tau, x_\tau \right\rangle + \psi\left(x_t\right) + \sup_{y \in X} \left\{ - \sum_{\tau=1}^{t} \left\langle g_\tau, y \right\rangle - t\psi(y) - \frac{t}{\eta} h_t(y) \right\} + h_t\left(x\right) \\
&= \frac{t}{\eta} \psi_t\left(x\right) + \sum_{\tau=1}^{t} \left\langle g_\tau, x_\tau \right\rangle + \psi\left(x_\tau\right) + \psi_t^*\left(-g_{1:t}\right)
\end{aligned}
$$

Proof of Proposition 2

Since $h_{t+1} \geq h_t$,

$$\begin{aligned}
\psi_t^* \left(-g_{1:t}\right) &= -\sum_{\tau=1}^{t} \left\langle g_\tau, x_{t+1} \right\rangle - t\psi\left(x_{t+1}\right) - \frac{1}{\eta} h_t\left(x_{t+1}\right) \\
&\leq -\sum_{\tau=1}^{t} \left\langle g_\tau, x_{t+1} \right\rangle - (t-1)\psi\left(x_{t+1}\right) - \psi\left(x_{t+1}\right) - \frac{1}{\eta}\psi_{t-1}\left(x_{t+1}\right) \\
&\leq \sup_{y \in \mathcal{X}} \left(- \left\langle g_{1:t}, y \right\rangle - (t-1)\psi(y) - \frac{1}{\eta} h_{t-1}(y) \right) - \varphi\left(x_{t+1}\right) \\
&= h_{t-1}^* \left(-g_{1:t}\right) - \psi\left(x_{t+1}\right)
\end{aligned}$$

Proof of Proposition 2

Identity (9) and the fact that $g_{1:t} - g_{1:t-1} = g_t$ give

$$
\begin{aligned}
R_t(x) &\leq \frac{1}{\eta} h_t(x) + \sum_{\tau=1}^{t} \langle g_\tau, x_\tau \rangle + \psi(x_{\tau+1}) + h_{t-1}^*(-g_{1:t}) - \psi(x_{t+1}) \\
&\leq \frac{1}{\eta} h_t(x) + \sum_{\tau=1}^{t} \langle g_\tau, x_\tau \rangle + \psi(x_{\tau+1}) - \psi(x_{t+1}) \\
&\quad + h_{t-1}^*(-g_{1:t-1}) - \langle \nabla h_{t-1}^*(g_{1:t-1}), g_t \rangle + \frac{\eta}{2} \|g_t\|_{h_{t-1}}^2 \\
&= \frac{1}{\eta} h_t(x) + \sum_{\tau=1}^{t} \langle g_\tau, x_\tau \rangle + \psi(x_{\tau+1}) + h_{t-1}^*(-g_{1:t-1}) + \frac{\eta}{2} \|g_t\|_{h_{t-1}}^2
\end{aligned}
$$

We can repeat the same steps that gaves the proposition 2.