Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization

Journal of Machine Learning Research (2010)

Lin Xiao

Presenter: Gyuseung Baek

June 3, 2019

Introduction

- Objective function(Loss function): $f_t(x) - f(x, z_t)$
  - f: loss function, z_t: data obtained at time t
- Regularzier: $\psi(x)$: closed convex function
- Let $\phi(x) = f(x) + \psi(x)$

- Online learning algorithm: from data $z_1, z_2, \cdots, z_t, \cdots$, achieve $x_1, x_2, \cdots, x_t, \cdots$
- Goal: Let $\phi^* = \min_x \phi(x)$. generate $\{x_t\}$ s.t.

$$\lim_{t \to \infty} \phi(x_t) = \phi^* \tag{1}$$

Stochastic Gradient Descent (SGD)

- SGD: Let $g_t \in \partial f_t(x_t)$. for some step size α_t,

$$x_{t+1} = x_t - \alpha_t(g_t + \xi_t)$$

  where ξ_t : subgradient of ψ at x_t
- If $\alpha_t = c/\sqrt{t}$, c: constant, $\{x_t\}$ satisties (1) with convergence rate $O(1/\sqrt{t})$
- It is indeed best possible for subgradient schemes with a *black-box* model (only function values and gradient informations are allowed) (Nemirovsky and Yudin, 1983)

- Drawbacks: At each time t, solution x_t does not satisfies the regularization

Regularized dual averaging (RDA)

- Update x_t as:

$$x_{t+1} = \arg\min_{x} \left\{ \frac{1}{t} \sum_{\tau=1}^{t} \langle g_\tau, w \rangle + \psi(x) + \frac{\beta_t}{t} h(x) \right\} \qquad (2)$$

  where $h(x)$ is an auxilaiary strongly convex function s.t. $\operatorname{argmin} h \subset \operatorname{argmin} \psi$, and $\{\beta_t\}_{t \geq 1}$ is a nonnegative and nondecreasing input sequence(learning rate).

- Convergence rates:
  - If $\beta_t = \Theta(\sqrt{t})$,

$$\mathbf{E}\phi\left(\overline{x}_t\right) - \phi^\star \leq O\left(\frac{G}{\sqrt{t}}\right)$$

    where $\overline{x}_t = (1/t)\sum_{\tau=1}^{t} x_\tau$, G: uniform upper bound on the norms of the subgradients g_t.
  - If ψ is strongly convex, then setting $\beta_t \leq O(lnt)$ gives a faster convergence rate $O(lnt/t)$
  - If $f(x, z)$ are all diff'ble and have Lipschitz continuous gradients (with const. L),

$$\mathbf{E}\phi\left(x_t\right) - \phi^\star \leq O(1)\left(\frac{L}{t^2} + \frac{Q}{\sqrt{t}}\right)$$

Regret bound

- Regret

$$R_t(x) \triangleq \sum_{\tau=1}^{t} \left(f_\tau\left(x_\tau\right) + \Psi\left(x_\tau\right)\right) - \sum_{\tau=1}^{t} \left(f_\tau(x) + \Psi(x)\right)$$

- If $\{x_t\}$ is acquired from simple SGD, $R_t(x) = O(\sqrt{t})$ for all $x \in dom\psi$.
- Similary, If $\{x_t\}$ is acquired from RDA,
  - $R_t(x) = O(\sqrt{t})$ for $\beta_t = \Theta(\sqrt{t})$
  - $R_t(x) = O(lnt)$ for $\beta_t = O(lnt), \psi$: strongly convex.

Strongly convex

- If a function f is convex, then

$$f(x) \geq f(y) + < g, x - y > + \frac{\sigma}{2}||x - y||^2, \quad \forall g \in \partial f(y) \qquad (3)$$

  for $\sigma \geq 0$.
- If there exist $\sigma > 0$ s.t. (3) hold for all $x \in dom f$, f is strongly convex with modulus σ w.r.t. norm $|| \cdot ||$.
- σ is called a convexity parameter of f.

Regret Bounds for online optimization

- Assumption
  - Suppose $h(x)$ is strongly convex with modulus 1 and $h(x) \leq D^2$ for all $x \in dom\psi$.
  - Let $\Gamma_D = \sup_{x:h(x)\leq D^2} \inf_{g\in\partial\psi(x)} \|g\|_*$
  - For all $t \geq 1$, there exist G s.t. $\|g_t\|_* \leq G$

- Let σ be a convexity parameter of ψ. If we set $\beta_0 = \max\{\sigma, \beta_1\}$, we can acquire the sequence of *regret bounds*

$$\Delta_t \triangleq \beta_t D^2 + \frac{G^2}{2}\sum_{\tau=0}^{t-1}\frac{1}{\sigma\tau+\beta_\tau} + \frac{2\left(\beta_0-\beta_1\right)G^2}{\left(\beta_1+\sigma\right)^2}, \quad t = 1, 2, 3, \ldots \tag{4}$$

Regret Bounds for online optimization

- **Thm 1** Let $\{x_t\}_{t \geq 1}$ be generated by RDA algorithm and above assumptions hold. Then for any $t \geq 1$, we have
  1. The regret bound is bounded by Δ_t
  $$R_t(x) \leq \Delta_t$$
  2. The primal variables are bounded as
  $$\|x_{t+1} - x\|^2 \leq \frac{2}{\sigma t + \beta_t} \left(\Delta_t - R_t(x)\right)$$
  3. If x in an interior point of *dom*ψ, then
  $$\|\overline{g_t}\|_* \leq \Gamma_D - \frac{1}{2}\sigma r + \frac{1}{rt} \left(\Delta_t - R_t(w)\right)$$

Regret Bounds for general convex regularization

- **Corr 1** Let $\{x_t\}_{t\geq 1}$ be generated by RDA algorithm and above assumptions hold. Set $\beta_t = \gamma\sqrt{t}$ for $t \geq 1$. Then for any $t \geq 1$, we have
  1. The regret bound is bounded as
  $$R_t(x) \leq \left(\gamma D^2 + \frac{G^2}{\gamma}\right)\sqrt{t}$$
  2. The primal variables are bounded as
  $$\|x_{t+1} - x\|^2 \leq D^2 + \frac{G^2}{\gamma^2} - \frac{1}{\gamma\sqrt{t}} R_t(x)$$
  3. If x in an interior point of *dom*ψ, then
  $$\|\overline{g_t}\|_* \leq \Gamma_D + \left(\gamma D^2 + \frac{G^2}{\gamma}\right)\frac{1}{r\sqrt{t}} - \frac{1}{rt} R_t(x)$$

- This bound is the same as the regret bound of SGD algorithm (Zinkevich, 2003)
- If we set $\gamma^\star = \frac{G}{D}$, then $R_t(x) \leq 2GD\sqrt{t}$

Regret Bounds for strongly convex regularization

- If ψ is strongly convex, $\beta_t \leq O(lnt)$ will give an $O(lnt)$ regret bound.
  - let $\beta_t = \sigma$ for $t \geq 0$. Then

$$\Delta_t = \sigma D^2 + \frac{G^2}{2\sigma} \sum_{t=0}^{t-1} \frac{1}{\tau+1} \leq \sigma D^2 + \frac{G^2}{2\sigma}(1 + \ln t)$$

  - let $\beta_t = \sigma(1 + \ln t)$ for $t \geq 1$. Then

$$\Delta_t = \sigma(1+\ln t)D^2 + \frac{G^2}{2\sigma}\left(1 + \sum_{\tau=1}^{t-1} \frac{1}{\tau + 1 + \ln \tau}\right) \leq \left(\sigma D^2 + \frac{G^2}{2\sigma}\right)(1+\ln t)$$

  - let $\beta_t = 0$ for $t \geq 1$. Then

$$\Delta_t = \frac{G^2}{2\sigma}\left(1 + \sum_{\tau=1}^{t-1} \frac{1}{\tau}\right) + \frac{2G^2}{\sigma} \leq \frac{G^2}{2\sigma}(6 + \ln t)$$

Convergence rates for stochastic learning

- **Thm 3** Assume $x^* = \operatorname{argmin}_x \phi(x)$ that satisifes $h(x^*) \leq D^2$ for some D and let $\phi^* = \phi(x^*)$. Let $\{x_t\}_{t\geq 1}$ be generated by RDA algorithm and assume $||g_t||_* \leq G$ for all $t \geq 1$. Then we have
  1. The expected cost associated with the random vairable $\bar{x}_t$ is bounded as
  $$\mathbf{E}\phi\left(\overline{x_t}\right) - \phi^\star \leq \frac{1}{t}\Delta_t$$
  2. The primal variables are bounded as
  $$\mathbf{E}\left\|x_{t+1} - x^\star\right\|^2 \leq \frac{2}{\sigma t + \beta_t}\Delta_t$$
  3. If x^* in an interior point of $dom\psi$, then
  $$\mathbf{E}\left\|\overline{g}_t\right\|_* \leq \Gamma_D - \frac{1}{2}\sigma r + \frac{1}{rt}\Delta_t$$

High probability bounds

- **Thm 5** Assume $h(x^*) \leq D^2$ for some D and for all $t \geq 1$, $h(x_t) \leq D^2$. Let $\{x_t\}_{t \geq 1}$ be generated by RDA algorithm and assume $||g_t||_* \leq G$ for all $t \geq 1$. Then for any $\delta \in (0, 1)$, we have, with probability at least $1 - \delta$,

$$\phi\left(\overline{x}_t\right) - \phi^\star \leq \frac{\Delta_t}{t} + \frac{8GD\sqrt{\ln(1/\delta)}}{\sqrt{t}} \tag{5}$$