Lecture 16 Online Newton Method

Recap: Follow the Regularized Leader

$$x_{t+1} = \min_{x \in K} \langle x, \sum_{t=1}^{t} \nabla_t \rangle + R(x)$$

$\nabla_t = \nabla f_t(x_t)$

Regret $\leq D_R G_R \sqrt{2T}$

$D_R = \sqrt{\max_{x,y \in K} (R(x) - R(y))}$

$G_R \geq \|\nabla_t\|_t^* \in \quad \forall t$

$\|\cdot\|_t = \|\cdot\|_{\nabla^2 R(z)}$

Recall key lemma (stability)

$$\text{Regret} \leq \sum_{t=1}^{T} \nabla_t^T (x_t - x_{t+1}) + \frac{1}{\eta} D_R^2$$

Long argument: $\nabla_t^T (x_t - x_{t+1}) \leq 2\eta \|\nabla_t\|_t^*$

1. we bounded

$$\sum_t \nabla_t^T x_t - \sum_t \nabla_t^T u$$

$$\text{Regret} \triangleq \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x^*),$$

why is this ok?

$$\rightarrow \sum_{t=1}^{T} f_t(x_t) - f_t(u) \leq \sum_{t=1}^{T} \nabla_t^T (x_t - u)$$

↑ Convex f_t

2. $R(x) = x^T \log x = \sum_{i=1}^{n} x_i \log x_i$ (negative entropy)

$K =$ Probability Simplex $= \Delta_n = \{x \mid x \geq 0, \sum x_i = 1\}$

$G_R = 1, \quad D_R \leq \log n$

$\Rightarrow \sqrt{T \log n}$ ←

Recovers the Hedge algorithm / exponentiated gradient

Alternate View from Strong Convexity

$\nabla_t^T (x_t - x_{t+1}) \leq \|\nabla_t\|_\infty \|x_t - x_{t+1}\|_1 = \|x_t - x_{t+1}\|_1$

$\leq \eta$

$\sum \nabla_t^T (x_t - x_n) \leq \eta T$

$R(x) = x^T \log x$ is 1-Strongly convex, in $\|\cdot\|_1$ ← HW4

for all $x \in \Delta_n$

$\|x_{t+1} - x_t\|_1 \leq ?$

$F(x_t) = \eta \langle \sum \nabla_t, x \rangle + R(x_t)$

$F(x_{t+1}) \geq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{1}{2}\|x_{t+1} - x_t\|_1^2$

$F(x_t) \geq F(x_{t+1}) + \langle \nabla F(x_{t+1}), x_t - x_{t+1} \rangle + \frac{1}{2}\|x_{t+1} - x_t\|_1^2$

$0 \geq \langle x_{t+1} - x_t, \nabla F(x_t) - \nabla F(x_{t+1}) \rangle + \|x_{t+1} - x_t\|_1^2$

$\|x_{t+1} - x_t\|_1^2 \leq \|x_{t+1} - x_t\|_1 \|\nabla F(x_t) - \nabla F(x_{t+1})\|_\infty \leq \eta$

$\|x_{t+1} - x_t\|_1 \leq \eta$

Universal Portfolio

$x_t \in \Delta_n$, asset allocation

$r_t \in R_+^n$ is price ratio $r_t(i) = \frac{Price_t(i)}{Price_{t-1}(i)}$

Wealth $W_{t+1} = W_t \cdot r_t^T x_t$

Total wealth $W_T = W_1 \cdot \prod_{t=1}^{T} r_t^T x_t$

after T days

$$\log \frac{W_T}{W_1} = \sum_{t=1}^{T} \log(r_t^T x_t) = -\sum_{t=1}^{T} f_t(x_t)$$

$f_t(x) = -\log(r_t^T x)$

$$Regret = \sum_{t=1}^{T} f_t(x_t) - \sum f_t(u)$$

$u \in \{e_1, e_2, \dots e_i \dots e_n\}$

NVDA AAPL TSLA

$u \in \Delta_n$

$r_t = \begin{pmatrix} 2 & \frac{1}{2} & 2 & \frac{1}{2} & 2 & \frac{1}{2} \cdots \\ \frac{1}{2} & 2 & \frac{1}{2} & 2 & \frac{1}{2} & 2 \cdots \end{pmatrix}$ $t=1$ $t=2$ $\cdots$ $t=T$

$u = [0.5, 0.5]$

$$W_T = W_1 \cdot \prod_{t=1}^{T} r_t^T u$$

$$= W_1 \cdot \left(\frac{1}{2} \cdot 2 + \frac{1}{2} \cdot \frac{1}{2}\right)^T$$

$$\geq W_1 \cdot 1.25^T$$

OGD: Regret $= G \cdot D\sqrt{T} = O(\sqrt{n}T)$

FTRL(entropy Regularizer): $O(\sqrt{T} \cdot \log n)$

Can we do better? improve from $\sqrt{T}$

when f_t is m-Strongly convex: OGD: $\frac{G^2}{m} \log T$

Is $f_t(x) = -\log(r_t^T x)$ strongly convex?

No.

Exponential Concavity:

A convex function f is α-exp-concave if $e^{-\alpha f(x)}$ is concave.

for all f s.t. f is twice differentiable

f is α-exp-concave at x

$$\iff \nabla^2 f(x) \geq \alpha \nabla f(x) \nabla f(x)^T$$

Recall Strong Convex:

$$\nabla^2 f(x) \geq m \cdot I$$

Example: $f(x) = -\log(r_t^T x)$

$$\nabla\left(-\log(r_t^T x)\right) = -\frac{r_t}{r_t^T x}$$

$$\nabla^2\left[-\log(r_t^T x)\right) = -\frac{0 - r_t \cdot r_t^T}{(r_t^T x)^2} = \frac{r_t r_t^T}{(r_t^T x)^2}$$

$$= \nabla \cdot \nabla^T$$

$$\nabla^2 \geq 1 \cdot \nabla \cdot \nabla^T \Rightarrow 1\text{-exp-concave.}$$

Example: f is m-Strongly Convex, in domain K s.t. $\|\nabla f\|_2 \leq G$

$$\nabla f \nabla f^T \preceq G^2 \cdot I \preceq \frac{G^2}{m} \nabla^2 f \quad \text{for all } x \in K$$

($\preceq$: G-Lipschitz; $\preceq$: m-Strongly Convex)

f is $\frac{m}{G^2}$-exp-concave.

$$\nabla^2 f \geq \frac{m}{G^2} \nabla f \nabla f^T$$

Useful Property of Exp-Concave Function

(Lemma 4.2 of OCO book Hazan)

f: α-exp-concave, domain K, G Lipschitz

$\forall x, y \in K$

$$f(x) \geq f(y) + \nabla f(y)^T (x-y) + \frac{\gamma}{2}(x-y)^T \nabla f(y) \nabla f(y)^T (x-y)$$

for $\gamma = \frac{1}{2}\min\{\alpha, \frac{1}{4GD}\}$

new for exp-concave functions

Alg: Online Newton Step

Input: K, T, $x_1 \in K$, $\gamma, \varepsilon > 0$, $A_0 = \varepsilon I$

for $t = 1, 2, \ldots, T$:

1. play x_t
2. incur loss $f_t(x_t)$, receive $\nabla_t = \nabla f_t(x_t)$
3. $A_t = A_{t-1} + \nabla_t \nabla_t^T$ quasi-Newton updates.

A_t: Newton step:

$$y_{t+1} = x_t - \frac{1}{\gamma} A_t^{-1} \nabla_t$$

(Projected)

$$x_{t+1} = \Pi_K^{A_t}(y_{t+1}) = \arg\min_{x \in K} \|y_{t+1} - x\|_{A_t}^2$$

Theorem Choose $\gamma = \frac{1}{2}\min\{\alpha, \frac{1}{4GD}\}$

$\varepsilon = \frac{1}{\gamma^2 D^2}$, for all $T \geq 4$

$$\text{Regret}_T \leq 5\left(\frac{1}{\alpha} + GD\right) n \log T$$

worse — Logarithmic in T

Lemma: $\text{Regret} \leq 4\left(\frac{1}{\alpha} + GD\right)\left(\sum_{t=1}^{T} \nabla_t^\top A_t^{-1} \nabla_t + 1\right)$

Proof: By the "Useful Property" of α-exp-concave

$$f_t(x^*) \geq f_t(x_t) + \nabla_t^\top (x^* - x_t) + \frac{\gamma}{2}(x^* - x_t)^\top \nabla_t \nabla_t^\top (x^* - x_t)$$

$$f_t(x_t) - f_t(x^*) \leq \nabla_t^\top (x_t - x^*) - \frac{\gamma}{2}(x^* - x_t)^\top \nabla_t \nabla_t^\top (x^* - x_t) \quad (*)$$

Now by the update rule of $x_{t+1} = \Pi_K^{A_t}\left(x_t - \frac{1}{\gamma} A_t^{-1} \nabla_t\right)$

$$\|x_{t+1} - x^*\|_{A_t}^2 \leq \|y_{t+1} - x^*\|_{A_t}^2 = \|x_t - \frac{1}{\gamma} A_t^{-1} \nabla_t - x^*\|_{A_t}^2$$

non-expansiveness in A_t of the projection

$$= \|x_t - x^*\|_{A_t}^2 + \frac{1}{\gamma^2}\|A_t^{-1} \nabla_t\|_{A_t}^2 - \frac{2}{\gamma}(x_t - x^*)^\top A_t \cdot A_t^{-1} \nabla_t$$

By move it around

$$\frac{2}{\gamma}(x_t - x^*)^T \nabla_t \le \frac{1}{\gamma^2}\|A_t^{-1}\nabla_t\|_{A_t}^2 + \|x_t - x^*\|_{A_t}^2 - \|x_{t+1} - x^*\|_{A_t}^2$$

multiply $\frac{\gamma}{2}$ on both sides and sum over $t = 1, 2, \dots, T$

$$\sum_{t=1}^{T}(x_t - x^*)^T \nabla_t \le \frac{1}{2\gamma}\sum_t \|\nabla_t\|_{A_t^{-1}}^2 + \frac{\gamma}{2}\|x_1 - x^*\|_{A_1}^2 + \frac{\gamma}{2}\sum_{t=2}^{T}(x_t - x^*)^T(A_t - A_{t-1})(x_t - x^*) - \frac{\gamma}{2}\|x_{T+1} - x^*\|_{A_T}^2$$

$A_t = A_{t-1} + \nabla_t \nabla_t^T$

$$\le \frac{1}{2\gamma}\sum_t \|\nabla_t\|_{A_t^{-1}}^2 + \frac{\gamma}{2}\|x_1 - x^*\|_{A_1}^2 + \sum_{t=2}^{T}(x_t - x^*)\nabla_t\nabla_t^T(x_t - x^*) + 0$$

dropping the negative term

plug into (*)

$$\sum_t f_t(x_t) - f_t(x^*) \le \frac{1}{2\gamma}\sum_t \|\nabla_t\|_{A_t^{-1}}^2 + \frac{\gamma}{2}\|x_1 - x^*\|_{A_1}^2 + \frac{\gamma}{2}\sum_{t=2}^{T}(x_t - x^*)\nabla_t\nabla_t^T(x_t - x^*)$$

$$\frac{\gamma}{2}\sum_{t=1}^{T}(x_t - x^*)\nabla_t\nabla_t^T(x_t - x^*)$$

$$= \frac{1}{2\gamma}\sum_t \|\nabla_t\|_{A_t^{-1}}^2 + \frac{\gamma}{2}(x_1 - x^*)^T A_1 (x_1 - x^*) - \frac{\gamma}{2}(x_1 - x^*)^T \nabla_t \nabla_t^T (x_1 - x^*)$$

$A_1 = \varepsilon I + \nabla_1 \nabla_1^T$

$$= \frac{1}{2\gamma}\sum_t \|\nabla_t\|_{A_t^{-1}}^2 + \frac{\gamma}{2}(x_1 - x^*)^T(A_1 - \nabla_t\nabla_t^T)(x_1 - x^*)$$

$$= \frac{1}{2\gamma}\sum_t \|\nabla_t\|_{A_t^{-1}}^2 + \frac{\gamma}{2}\varepsilon\|x_1 - x^*\|_2^2 \qquad \varepsilon = \frac{1}{\gamma^2 D^2}$$

$$\le \frac{\gamma}{2}\varepsilon \cdot D^2 = \frac{1}{2\gamma}$$

$$= \frac{1}{2\gamma}\left(\sum_t \|\nabla_t\|_{A_t^{-1}}^2 + 1\right)$$

$\gamma = \frac{1}{2}\min\left(\alpha, \frac{1}{40}\right)$

$$\le 4\left(\frac{1}{\alpha} + GD\right)\left(\sum_t \|\nabla_t\|_{A_t^{-1}}^2 + 1\right)$$

□

$$\Sigma f_t(x_t) - f_t(x^*) \le 4\left(\frac{1}{\alpha} + GD\right)\left(\sum_{t=1}^{T} \|\nabla_t\|_{A_t^{-1}}^2 + 1\right)$$

Remains to bound this term!

Proof of the main Theorem:

$$\|\nabla_t\|_{A_t^{-1}}^2 = \nabla_t^\top (A_t)^{-1} \nabla_t = \mathrm{tr}\left(\nabla_t^\top A_t^{-1} \nabla_t\right) = \mathrm{tr}\left(\underbrace{A_t^{-1}}_{\mathbb{R}^{n\times n}} \underbrace{\nabla_t \nabla_t^\top}_{\mathbb{R}^{n\times n}}\right) =$$

$$A_t^{-1} \bullet \left(\nabla_t \nabla_t^\top\right).$$

$$= \mathrm{tr}\left(A_t^{-1}(A_t - A_{t-1})\right) = \mathrm{tr}\left(I - A_t^{-1} A_{t-1}\right) = \sum_i \left(1 - \lambda_i(A_t^{-1} A_{t-1})\right)$$

$$\underbrace{\xi \le \log\left(\frac{1}{1-\xi}\right)}_{\text{take}} \le \sum_{i=1}^{n} \log\left(\lambda_i^{-1}(A_t^{-1} A_{t-1})\right) = \log\left(\prod_{i=1}^{n} \lambda_i^{-1}(A_t^{-1} A_{t-1})\right) = \log\left|(A_t^{-1} A_{t-1})^{-1}\right|$$

take $\xi = 1 - \lambda_i(A_t^{-1} A_{t-1})$

$$= \log\left(\frac{|A_t|}{|A_{t-1}|}\right) = \log|A_t| - \log|A_{t-1}|$$

add up $t = 1, 2, \cdots, T$

$$\sum_{t=1}^{T} \|\nabla_t\|_{A_t^{-1}}^2 = \log|A_T| - \log|A_0|$$

$A_0 = \varepsilon I$, $|A_0| = \varepsilon^n$

$A_T = \sum_{t=1}^{T} \nabla_t \nabla_t^\top + \varepsilon I \preceq (TG^2 + \varepsilon) I$

$|A_T| \le (TG^2 + \varepsilon)^n$

$\gamma = \frac{1}{2}\min\left(\alpha, \frac{1}{4GD}\right) \le n \cdot \log\left(\frac{TG^2}{\varepsilon} + 1\right)$

$\xi = \frac{1}{\gamma^2 D^2} = n \cdot \log\left(\gamma^2 T G^2 D^2 + 1\right) \le n \cdot \log\left(\frac{1}{64} T \frac{G^2 D^2}{G^2 D^2} + 1\right) \simeq n \cdot \log T$ for $T \ge 4$. □

Complexity (computational)

$A_t^{-1} v_t$ requires $O(n^3)$

space $O(n^2)$

Sherman-Morrison-Woodbury (Matrix inversion lemma)

$$(A + xx^T)^{-1} = A^{-1} - \frac{A^{-1} x x^T A^{-1}}{1 + x^T A^{-1} x}$$

A^{-1}: Newton step from previous iteration

$O(n^2)$

Projection $\min_{x \in K} \|y_{t+1} - x\|_{A_t}^2$ $O(n^2 \log \frac{1}{\varepsilon})$

$\downarrow$

$O(n^2)$

OGD: $O(n)$ space and time