**EE 381V: Large Scale Optimization** **Fall 2012**

Lecture 24 — November 27

*Lecturer: Caramanis & Sanghavi* *Scribe: Jahshan Bhatti and Ken Pesyna*

# 24.1 Last Time: Mirror Descent

The convergence of subgradient descent is given by

$$f\left(x^\star_{\text{best}}\right) = f^\star \le \frac{L \cdot R}{\sqrt{k+1}} \tag{24.1}$$

where $L$ is the Lipschitz constant with respect to $\|\cdot\|_2$ and $R$ is the size of the set $\|x_0 - x^\star\|_2$. The subgradient update is given by

$$\begin{aligned} x^+ &= \text{Proj}_{\mathscr{X}}\left(x - \gamma_t g\right) &(24.2)\\ &= \arg\min_{u \in \mathscr{X}} \left[\left\langle \gamma g - \nabla w\left(x\right), u\right\rangle + w\left(u\right)\right] &(24.3)\end{aligned}$$

where $g \in \partial f\left(x\right)$ and $w\left(u\right) = \frac{1}{2}\|u\|_2^2$ is the "distance generating function" that is continuous, differentiable, and strongly convex with respect to $\|\cdot\|_2$. The idea is to replace $w$ with some other function. The bounds are replaced by $L \to L_f$ and $R \to$ "size of set" measured by Bregman divergence given by DGF $w\left(\cdot\right)$. Also, $w\left(\cdot\right)$ is $\alpha$-strongly convex with respect to $\|\cdot\|$.

# 24.2 Analysis of Convergence

In Euclidean case: Guaranteed decrease in Lyapunov function ($\|x_k - x^\star\|_2$). For any $u \in \mathscr{X}$,

$$\frac{1}{2}\|x-u\|_2^2 - \frac{1}{2}\|x_+ - u\|_2^2 \ge \gamma\left\langle g, x-u\right\rangle - \frac{1}{2}\gamma^2\|g\|_2^2 \tag{24.4}$$

The Bregman Divergence of $\|u-v\|_2^2$ is given by

$$D\left(u,v\right) = w\left(u\right) - w\left(v\right) - \left\langle \nabla w\left(v\right), u-v\right\rangle \tag{24.5}$$

Analog to key inequality:

$$D\left(u, x_t\right) - D\left(u, x_{t+1}\right) \ge \gamma_t\left\langle g_t, x_t - u\right\rangle - \frac{1}{2\alpha}\gamma_t^2\|g_t\|_\star^2 \tag{24.6}$$

$$\underbrace{\left[\left\langle \nabla w\left(x_t\right), x_t - u\right\rangle - w\left(x_t\right)\right]}_{H_u(x_t)} - \underbrace{\left[\left\langle \nabla w\left(x_{t+1}\right), x_{t+1} - u\right\rangle - w\left(x_{t+1}\right)\right]}_{H_u(x_{t+1})}$$
$$\geq \gamma_t \left\langle g_t, x_t - u\right\rangle - \frac{1}{2\alpha}\sum \gamma_t^2 \left\|g_t\right\|_\star^2 \quad (24.7)$$

Recall

$$f\left(u\right) \geq f\left(x_t\right) + \left\langle g_t, u - x_t\right\rangle \quad (24.8)$$
$$\gamma_t\left(f\left(x_t\right) - f\left(u\right)\right) \leq \gamma_t\left\langle g_t, x_t - u\right\rangle \quad (24.9)$$

Summing (24.7) from $t = 0$ to $t = T$ yields

$$\sum_{t=0}^{T} \gamma_t \left\langle g_t, x_t - u\right\rangle \leq \underbrace{H_u\left(x_0\right) - H_u\left(x_T\right)}_{\Theta} + \frac{1}{2\alpha}\sum \gamma_t^2 \left\|g_t\right\|_\star^2 \quad (24.10)$$
$$\sum \gamma_t \underbrace{\left(f\left(x_t\right) - f\left(u\right)\right)}_{f\left(x_{\text{best}}^T\right) \leq f\left(x_t\right)} \leq \quad (24.11)$$
$$\underbrace{\left(f\left(x_{\text{best}}^T\right) - f\left(u\right)\right)}_{\text{Let } u = x^\star} \sum \gamma_t \leq \Theta + \frac{1}{2\alpha}\sum \gamma_t^2 \left\|g_t\right\|_\star^2 \quad (24.12)$$
$$f\left(x_{\text{best}}^T\right) - f^\star \leq \frac{\Theta + \frac{1}{2\alpha}\sum \gamma_t^2 \left\|g_t\right\|_\star^2}{\sum \gamma_t} \quad (24.13)$$

where $\Theta$ is the upper bound on $\left\|x^\star - x_0\right\|_2^2 = \operatorname{diam}\mathscr{X}$ or generally "size of $\mathscr{X}$ measured by $D\left(\cdot,\cdot\right)$."

Take

$$\gamma_t = \frac{\sqrt{\Theta \cdot \alpha}}{\left\|g_t\right\|_\star \cdot \sqrt{t}} \quad (24.14)$$

Exercise:

$$\epsilon_T \quad \leq \quad O\left(1\right)\frac{\sqrt{\Theta}L_{\|\cdot\|}^F}{\sqrt{2}\sqrt{T}}.If \left\|\cdot\right\| = \left\|\cdot\right\|_2, w = \frac{1}{2}\left\|\cdot\right\|_2$$

For

$$X \quad \in \quad \Delta_n^+(R),\ w\left(x\right) = \sum x_i \ln\left(x_i\right), \left\|\cdot\right\| = \left\|\cdot\right\|_2$$

then mirror descent update is easy

Exercise:

$$\begin{aligned} \alpha &= O\left(1\right)/R^2 (\text{modulus of strong convexity w.r.t. } \left\|\cdot\right\|_1 \\ \Theta &\leq O\left(1\right)\ln\left(n\right) \\ \epsilon_T &\leq O\left(1\right)\sqrt{\ln\left(n\right)}\frac{L_{\|\cdot\|}^F R}{\sqrt{T}} \end{aligned}$$

Mirror Descent versus Subgradient Descent (Error Ratio)

$$\frac{\epsilon_{MD}}{\epsilon_{SD}} = \underbrace{O\left(\sqrt{\ln(n)}\right)}_{(I)} \cdot \underbrace{\frac{\max_X \|x-y\|_1}{\max_X \|x-y\|_2}}_{(II)} \cdot \underbrace{\frac{L^f_{\|\cdot\|_1}}{L^f_{\|\cdot\|_2}}}_{(III)}$$

Analysis:

- (I) Always Favors Euclidean
- (II) Always favors Euclidean ($1 \leq ratio \leq \sqrt{n}$)
- (III) Favors MD-simplex ($\frac{1}{\sqrt{n}} \leq ratio \leq 1$)
- For $\mathscr{X}$ ball, $f$ is sensitive to $O(1)$ coordinate $\rightarrow$ subgradient descent much better: $\sqrt{n\ln(n)}$
- For $\mathscr{X}$ simplex, $f$ is sensitive to $O(n)$ coordinates $\rightarrow$ MD-Simplex better: $\frac{\sqrt{n}}{\sqrt{\ln(n)}}$

## 24.3 Algorithms that use the Dual

Recall Duality

Primal:

$$\min_x f(x) \; s.t. \; h(x) \quad \leq \quad 0, Ax = b$$

Lagrangian:

$$\begin{aligned} \mathscr{L}_{\lambda\geq 0}(x,\lambda,\nu) &= f(x) + \lambda^T h(x) + \nu(Ax-b) \\ g(\lambda,\nu) &= \min_x \mathscr{L}(x,\lambda,\nu) \end{aligned}$$

Dual:

$$\lambda^\star, \nu^\star \quad = \quad arg \max_{\lambda\geq 0,\nu} g(\lambda,\nu)$$

Then can get primal back by

$$x^\star = arg\min_x \mathscr{L}(x,\lambda^\star,\nu^\star)$$

### 24.3.1 Primal and Dual Decomposition

Idea: Use the problem structure for faster/parallel solution

- Complicating variable
- Complicating constraint

Complicating Variable:

$$
\begin{array}{rll}
\text{subproblem 1} & \min_{x_1} & f_1(x,y)\}\phi_1(y) \\
\text{subproblem 2} & \min_{x_2} & f_2(x,y)\}\phi_2(y) \\
\text{master problem} & \min_{y} & \phi_1(y)+\phi_2(y)
\end{array}
$$

Options to solve: Bisection, take gradient of $\phi$, solve $\phi_1, \phi_2$ exactly $\rightarrow$ Doesn't matter

### 24.3.2 Dual Decomposition

$$
\begin{array}{rcl}
\min\limits_{x_1 y_1 x_2 y_2} & & f_1(x_1,y_1)+f_2(x_2,y_2)\ s.t.\ y_1=y_2 \\
\mathscr{L}(x_1,y_1,x_2,y_2) & = & f_1(\cdot)+f_2(\cdot)+\lambda(y_1-y_2) \\
\text{subproblem 1} & & \min\limits_{x_1 y_1} f_1(x_1,y_1)+\lambda y_1 \\
\text{subproblem 2} & & \min\limits_{x_2 y_2} f_2(x_2,y_2)-\lambda y_2 \\
\lambda_+ & = & \lambda-\alpha(y_2-y_1)
\end{array}
$$