24.1 Last Time: Mirror Descent

The convergence of subgradient descent is given by

\[ f(x^{\star}_{\text{best}}) = f^{\star} \leq \frac{L \cdot R}{\sqrt{k+1}} \] (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

\[ x^+ = \text{Proj}_{\mathcal{X}}(x - \gamma_t g) \] (24.2)

\[ = \arg \min_{u \in \mathcal{X}} [\langle \gamma g - \nabla w(x), u \rangle + w(u)] \] (24.3)

where \( g \in \partial f(x) \) and \( w(u) = \frac{1}{2} \|u\|_2^2 \) 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 \( R \to “\text{size of set}” \) measured by Bregman divergence given by DGF \( w(\cdot) \). Also, \( w(\cdot) \) 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 \( u \in \mathcal{X}^\star \),

\[ \frac{1}{2} \|x - u\|_2^2 - \frac{1}{2} \|x_+ - u\|_2^2 \geq \gamma \langle g, x - u \rangle - \frac{1}{2} \gamma^2 \|g\|_2^2 \] (24.4)

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

\[ D(u,v) = w(u) - w(v) - \langle \nabla w(v), u - v \rangle \] (24.5)

Analog to key inequality:

\[ D(u,x_t) - D(u,x_{t+1}) \geq \gamma_t \langle g_t, x_t - u \rangle - \frac{1}{2\alpha} \gamma_t^2 \|g_t\|_2^2 \] (24.6)
\[
\left[ \langle \nabla w(x_t), x_t - u \rangle - w(x_t) \right]_{H_u(x_t)} - \left[ \langle \nabla w(x_{t+1}), x_{t+1} - u \rangle - w(x_{t+1}) \right]_{H_u(x_{t+1})} \\
\geq \gamma_t \langle g_t, x_t - u \rangle - \frac{1}{2\alpha} \sum \gamma_t^2 \| g_t \|^2 \tag{24.7}
\]

Recall

\[
f(u) \geq f(x_t) + \langle g_t, u - x_t \rangle \tag{24.8}
\]

\[
\gamma_t (f(x_t) - f(u)) \leq \gamma_t \langle g_t, x_t - u \rangle \tag{24.9}
\]

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

\[
\sum_{t=0}^{T} \gamma_t \langle g_t, x_t - u \rangle \leq H_u(x_0) - H_u(x_T) + \frac{1}{2\alpha} \sum \gamma_t^2 \| g_t \|^2 \tag{24.10}
\]

\[
\sum_{f(x_t)} \gamma_t (f(x_t) - f(u)) \leq \frac{1}{2\alpha} \sum \gamma_t^2 \| g_t \|^2 \tag{24.11}
\]

Let \( u = x^* \)

\[
(f(x_T) - f(u)) \sum \gamma_t \leq \Theta + \frac{1}{2\alpha} \sum \gamma_t^2 \| g_t \|^2 \tag{24.12}
\]

\[
f(x_T) - f^* \leq \frac{\Theta}{\sum \gamma_t} + \frac{1}{2\alpha} \sum \gamma_t^2 \| g_t \|^2 \tag{24.13}
\]

where \( \Theta \) is the upper bound on \( \| x^* - x_0 \|^2 = \text{diam} \mathcal{X} \) or generally “size of \( \mathcal{X} \)” measured by \( D(\cdot, \cdot) \).

Take

\[
\gamma_t = \frac{\sqrt{\Theta} \cdot \alpha}{\| g_t \|^2 \cdot \sqrt{t}} \tag{24.14}
\]

**Exercise:**

\[
\epsilon_T \leq O\left(1\right) \frac{\sqrt{\Theta} L_n^F}{\sqrt{2\sqrt{T}}} I f \| \cdot \| = \| \cdot \|_2, w = \frac{1}{2} \| \cdot \|_2
\]

For

\[
X \in \Delta_n^+(R), \ w(x) = \sum x_i \ln(x_i), \| \cdot \| = \| \cdot \|_2
\]

then mirror descent update is easy

**Exercise:**

\[
\alpha = O\left(1\right) / R^2 (\text{modulus of strong convexity w.r.t. } \| \cdot \|_1)
\]

\[
\Theta \leq O\left(1\right) \ln(n)
\]

\[
\epsilon_T \leq O\left(1\right) \frac{\sqrt{\ln(n)} L_n^F R}{\sqrt{T}}
\]
Mirror Descent versus Subgradient Descent (Error Ratio)

\[ \frac{\epsilon_{MD}}{\epsilon_{SD}} = O\left(\sqrt{\ln(n)}\right) \cdot \frac{\max_X \| x - y \|_1}{\max_X \| x - y \|_2} \cdot \frac{L_f}{\| x \|_2} \]

Analysis:

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

### 24.3 Algorithms that use the Dual

Recall Duality

**Primal:**

\[
\min_x f(x) \text{ s.t. } h(x) \leq 0, Ax = b
\]

**Lagrangian:**

\[
\mathcal{L}_{\lambda \geq 0}(x, \lambda, \nu) = f(x) + \lambda^T h(x) + \nu(Ax - b)
\]

\[
g(\lambda, \nu) = \min_x \mathcal{L}(x, \lambda, \nu)
\]

**Dual:**

\[
\lambda^*, \nu^* = \arg\max_{\lambda \geq 0, \nu} g(\lambda, \nu)
\]

Then can get primal back by

\[
x^* = \arg\min_x \mathcal{L}(x, \lambda^*, \nu^*)
\]
24.3.1 Primal and Dual Decomposition

Idea: Use the problem structure for faster/parallel solution

- Complicating variable
- Complicating constraint

Complicating Variable:

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

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

24.3.2 Dual Decomposition

\begin{align*}
\min_{x_1, y_1, x_2, y_2} & \quad f_1(x_1, y_1) + f_2(x_2, y_2) \quad \text{s.t.} \quad y_1 = y_2 \\
\mathcal{L}(x_1, y_1, x_2, y_2) & = f_1(\cdot) + f_2(\cdot) + \lambda (y_1 - y_2) \\
\text{subproblem 1} & \quad \min_{x_1, y_1} f_1(x_1, y_1) + \lambda y_1 \\
\text{subproblem 2} & \quad \min_{x_2, y_2} f_2(x_2, y_2) - \lambda y_2 \\
\lambda_+ & = \lambda - \alpha (y_2 - y_1)
\end{align*}