MATH466/MATH766
Math of machine learning
Appendix Convergence Analysis

References:

•

Todays contents: convergence analysis of

- gradient descent
- subgradient method
- proximal gradient
- stochastic gradient

Important concepts:

•

Recommend reading:

•

GD

$$\min_{x \in \mathbb{R}^n} f(x) \qquad x^{(k+1)} = x^{(k)} - t \nabla f(x^{(k)})$$

Thm 1. Assume f is L-smooth, and $f(x) \geq f^*$ for any x

then with $t = \frac{1}{L}$, $\min \{ \| \nabla f(x^{(k)}) \|^2 \}_{k=0}^{k-1} \leq \frac{2L (f(x^{(0)}) - f^*)}{K}$

Thm 2. In addition to Thm 1 assumptions,

if f is convex and $f(x) \geq f(x^*)$ for any x.

then $f(x^{(k)}) - f(x^*) \leq \frac{L \| x^{(0)} - x^* \|^2}{2k}$

Thm 3. In addition to Thm 1, 2 assumptions,

if f is μ-strongly convex.

then $f(x^{(k)}) - f(x^*) \leq (1 - \frac{\mu}{L})^k (f(x^{(0)}) - f(x^*))$

and $\| x^{(k)} - x^* \|^2 \leq (1 - \frac{\mu}{L})^k \| x^{(0)} - x^* \|^2$

Proof. 1. By Lipschitz smoothness,

$$f(x^{(k+1)}) \leq f(x^{(k)}) - t \| \nabla f(x^{(k)}) \|_2^2 + \frac{Lt^2}{2} \| \nabla f(x^{(k)}) \|_2^2$$

$$= f(x^{(k)}) - t\left(1 - \frac{Lt}{2}\right) \| \nabla f(x^{(k)}) \|_2^2$$

Pick $t = \frac{1}{L}$

$$f(x^{(k+1)}) \leq f(x^{(k)}) - \frac{1}{2L} \| \nabla f(x^{(k)}) \|_2^2 \qquad (*)$$

$$\Rightarrow \quad \| \nabla f(x^{(k)}) \|_2^2 \leq 2L (f(x^{(k)}) - f(x^{(k+1)}))$$

$$\Rightarrow \quad \sum_{k=0}^{k-1} \| \nabla f(x^{(k)}) \|_2^2 \leq 2L (f(x^{(0)}) - f(x^{(k)}))$$

$$\leq 2L (f(x^{(0)}) - f^*) < \infty$$

therefore $\min \{ \| \nabla f(x^{(k)}) \|_2^2 \}_{k=0}^{k-1} \leq \frac{2L (f(x^{(0)}) - f^*)}{K}$

2. From L-smooth, we have

$$f(x^{(k+1)}) \le f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \rangle + \frac{L}{2} \| x^{(k+1)} - x^{(k)} \|^2$$

By μ-convexity, we have

$$f(x^{(k)}) + \langle \nabla f(x^{(k)}), x^* - x^{(k)} \rangle + \frac{\mu}{2} \| x^{(k)} - x^* \|^2 \le f(x^*)$$

Combine above two, we have

$$f(x^{(k+1)}) - f(x^*) \le \langle \nabla f(x^{(k)}), x^{(k+1)} - x^* \rangle + \frac{L}{2} \| x^{(k+1)} - x^{(k)} \|^2 - \frac{\mu}{2} \| x^{(k)} - x^* \|^2 \quad (**)$$

Let $a := x^{(k)} - x^*$, $b := x^{(k+1)} - x^*$,

then $x^{(k+1)} - x^{(k)} = b - a$, $\nabla f(x^{(k)}) = -\frac{1}{t}(b - a)$

RHS of $(**) = -\frac{1}{t} \langle b-a, b \rangle + \frac{L}{2} \| b-a \|^2 - \frac{\mu}{2} \| a \|^2$

$$\overset{t=\frac{1}{L}}{=\!=} L \langle b-a, -b + \frac{1}{2}(b-a) \rangle$$

$$= -\frac{L}{2} \langle b-a, b+a \rangle$$

$$= \frac{L-\mu}{2} \| x^{(k)} - x^* \|^2 - \frac{L}{2} \| x^{(k+1)} - x^* \|^2 \quad (\Delta)$$

When $\mu = 0$

$$f(x^{(k+1)}) - f(x^*) \le \frac{L}{2} \left(\| x^{(k)} - x^* \|^2 - \| x^{(k+1)} - x^* \|^2 \right)$$

$$\Rightarrow \sum_{k=0}^{K} \left(f(x^{(k)}) - f(x^*) \right) \le \frac{L}{2} \left(\| x^{(0)} - x^* \|^2 - \| x^{(k+1)} - x^* \|^2 \right)$$

$$\le \frac{L}{2} \| x^{(0)} - x^* \|^2$$

Since $f(x^{(k+1)}) \le f(x^{(k)})$

$$f(x^{(K)}) - f(x^*) \le \frac{L \| x^{(0)} - x^* \|^2}{2K}$$

3. From 1, we have

$$f(x^{(k+1)}) - f(x^*) \le f(x^{(k)}) - f(x^*) - t\left(1 - \frac{Lt}{2}\right)\|\nabla f(x^{(k)})\|^2 \qquad (*)$$

Intuition: strong convexity.

$\Rightarrow$ if $\|\nabla f(x)\|$ is small ($\|\nabla f(x)\|$ close to $\|\nabla f(x^*)\|$)

then x is close to x^*, $f(x)$ is close to $f(x^*)$

By strong convexity

$$h(y) := \underbrace{f(x^{(k)}) + \langle \nabla f(x^{(k)}), y - x^{(k)} \rangle + \frac{\mu}{2}\|y - x^{(k)}\|^2}_{\text{quadratic in } y} \le f(y), \quad \forall y$$

$$h(y) \ge h\left(x^{(k)} - \frac{1}{\mu}\nabla f(x^{(k)})\right) = f(x^{(k)}) - \frac{1}{2\mu}\|\nabla f(x^{(k)})\|^2$$

$$\Rightarrow \quad -\|\nabla f(x^{(k)})\|^2 \le 2\mu\left(f(x^*) - f(x^{(k)})\right)$$

Plug in $(*)$ $\quad f(x^{(k+1)}) - f(x^*) \le \left[1 - 2\mu t\left(1 - \frac{Lt}{2}\right)\right]\left(f(x^{(k)}) - f(x^*)\right)$

$$= \left(1 - 2\mu t + \mu L t^2\right)\left(f(x^{(k)}) - f(x^*)\right) \qquad (***)$$

Let $C(t) := 1 - 2\mu t + \mu L t^2 = \mu L\left(t - \frac{1}{L}\right)^2 + \left(1 - \frac{\mu}{L}\right)$

When $t \in \left(0, \frac{2}{L}\right)$, $C(t) \in (0, 1)$.

For largest shrinkage, take $t = \frac{1}{L}$

$$f(x^{(k+1)}) - f(x^*) \le \left(1 - \frac{\mu}{L}\right)\left(f(x^{(k)}) - f(x^*)\right)$$

By induction $\quad f(x^{(k)}) - f(x^*) \le \left(1 - \frac{\mu}{L}\right)^k \left(f(x^{(0)}) - f(x^*)\right)$

From 2 (Δ)

$$0 \le f(x^{(k+1)}) - f(x^*) \le \frac{L - \mu}{2}\|x^{(k)} - x^*\|^2 - \frac{L}{2}\|x^{(k+1)} - x^*\|^2$$

$$\Rightarrow \quad \|x^{(k+1)} - x^*\|^2 \le \left(1 - \frac{\mu}{L}\right)\|x^{(k)} - x^*\|^2$$

$$\Rightarrow \quad \|x^{(k)} - x^*\|^2 \le \left(1 - \frac{\mu}{L}\right)^k \|x^{(0)} - x^*\|^2$$

SubGrad

$$\min_{x \in \mathbb{R}^n} f(x) \qquad x^{(k+1)} = x^{(k)} - t_k g_k, \quad g_k \in \partial f(x^{(k)})$$

$$f(x^{(k)}_{best}) := \min_{k=0,\dots,K} f(x^{(k)})$$

Thm. Assume that f is G-Lipschitz continuous, i.e. $|f(x) - f(y)| \le G\|x-y\|_2$

Then $$f(x^{(k)}_{best}) - f(x^*) \le \frac{\|x^{(1)} - x^*\|^2 + G^2 \sum_{k=1}^{K} t_k^2}{2 \sum_{k=1}^{K} t_k}$$

Pf. By def of subgrad

$$f(x^*) \ge f(x^{(k)}) + \langle g_k, x^* - x^{(k)} \rangle$$

$$\Rightarrow f(x^{(k)}) - f(x^*) \le \langle g_k, x^* - x^{(k)} \rangle$$

$$= \frac{1}{t_k} \langle x^{(k)} - x^{(k+1)}, x^{(k)} - x^* \rangle \qquad (*)$$

Let $a := x^{(k)} - x^*$, $b := x^{(k+1)} - x^*$,

then $b - a = x^{(k+1)} - x^{(k)} = -t_k g_k$

RHS of (*) $= \frac{1}{t_k} \langle a - b, \frac{a+b}{2} + \frac{a-b}{2} \rangle$

$$= \frac{1}{2t_k} \left(\|a\|^2 - \|b\|^2 \right) + \frac{t_k}{2} \|g_k\|^2$$

$$\Rightarrow 2t_k \left(f(x^{(k)}) - f(x^*) \right) \le \|x^{(k)} - x^*\|^2 - \|x^{(k+1)} - x^*\|^2 + t_k^2 \|g_k\|^2$$

$$\Rightarrow \sum_{k=1}^{K} 2t_k \left(f(x^{(k)}) - f(x^*) \right) \le \|x^{(1)} - x^*\|^2 + \sum_{k=1}^{K} t_k^2 \|g_k\|^2$$

$$\Rightarrow f(x^{(k)}_{best}) - f(x^*) \le \frac{\|x^{(1)} - x^*\|^2 + \sum_{k=1}^{K} t_k^2 \|g_k\|^2}{2 \sum_{k=1}^{K} t_k}$$

ProxGD $\min_{x \in \mathbb{R}^n} f(x) = h(x) + r(x)$

h differentiable, $\quad x^{(k+1)} = \text{prox}_{tr}(x^{(k)} - t \nabla h(x^{(k)}))$

r not necessarily differentiable

Thm. If h is convex, differentiable and L-smooth,

r is convex and $\text{prox}_{tr}(x)$ can be evaluated

then proximal gradient descent with fixed step size $t \le \frac{1}{L}$

satisfies $\quad f(x^{(k)}) - f^* \le \frac{\|x^{(0)} - x^*\|_2^2}{2tk}$

Proof. Denote $\quad G_t(x) := \frac{1}{t}(x - \text{prox}_{tr}(x - t \nabla h(x)))$

Then $\quad x^{(k+1)} = x^{(k)} - t\, G_t(x^{(k)})$

(Central Lemma)

Claim: For any z,

$$f(x^{(k+1)}) \le f(z) + \langle G_t(x^{(k)}), x^{(k)} - z \rangle - (t - \frac{Lt^2}{2}) \|G_t(x^{(k)})\|^2$$

Take $t = \frac{1}{L}$

Take $z = x^{(k)} \Rightarrow f(x^{(k+1)}) \le f(x^{(k)}) - \underbrace{\frac{t}{2}\|G_t(x^{(k)})\|^2}_{\ge 0}$

make progress each iter

Take $z = x^*$

$$\Rightarrow f(x^{(k+1)}) - f(x^*) \le \frac{1}{t}\langle x^{(k)} - x^{(k+1)}, x^{(k)} - x^* \rangle - \frac{1}{2t}\|x^{(k)} - x^{(k+1)}\|^2$$

similar strategy

$$f(x^{(k+1)}) - f(x^*) \le \frac{1}{2t}\left(\|x^{(k)} - x^*\|^2 - \|x^{(k+1)} - x^*\|^2\right)$$

$$\Rightarrow f(x^{(k)}) - f(x^*) \le \frac{\|x^{(0)} - x^*\|^2}{2tk}$$

Proof of Claim:

(Differentiable part)

By L-smooth, we have

$$h(x^{(k+1)}) \le h(x^{(k)}) + \langle \nabla h(x^{(k)}),\, x^{(k+1)} - x^{(k)} \rangle + \frac{L}{2} \| x^{(k+1)} - x^{(k)} \|^2$$

By convexity of h, we have

$$h(x^{(k)}) + \langle \nabla h(x^{(k)}),\, z - x^{(k)} \rangle \le h(z)$$

Combine above

$$\Rightarrow h(x^{(k+1)}) \le h(z) + \langle \nabla h(x^{(k)}),\, x^{(k)} - z \rangle + \langle \nabla h(x^{(k)}),\, x^{(k+1)} - x^{(k)} \rangle + \frac{L}{2} \| x^{(k+1)} - x^{(k)} \|^2 \quad (*)$$

(Non-differentiable part)

By convexity of r, for any $g \in \partial r(w^{(k+1)})$, z

$$r(x^{(k+1)}) + \langle g,\, z - x^{(k+1)} \rangle \le r(z) \quad (**)$$

By update rule

$$x^{(k+1)} = \operatorname*{argmin}_x \; r(x) + \frac{1}{2t} \| x - (x^{(k)} - t \nabla h(x^{(k)})) \|^2$$

By optimality condition

$$0 \in \partial r(x^{(k+1)}) + \frac{1}{t} (x^{(k+1)} - x^{(k)} + t \nabla h(x^{(k)}))$$

$$\Rightarrow \frac{1}{t} (x^{(k)} - x^{(k+1)}) - \nabla h(x^{(k)}) \in \partial r(x^{(k+1)})$$

$$\Rightarrow G_t(x^{(k)}) - \nabla h(x^{(k)}) \in \partial r(x^{(k+1)})$$

Combine with (**)

$$r(x^{(k+1)}) \le r(z) + \langle \nabla h(x^{(k)}) - G_t(x^{(k)}),\, z - x^{(k+1)} \rangle \quad , \forall z \quad (***)$$

(*) + (***)

$$f(x^{(k+1)}) \le f(z) + \langle G_t(x^{(k)}),\, x^{(k+1)} - z \rangle + \frac{L}{2} \| x^{(k+1)} - x^{(k)} \|^2$$

$$= f(z) + \langle G_t(x^{(k)}),\, x^{(k)} - z \rangle - (t - \frac{Lt^2}{2}) \| G_t(x^{(k)}) \|^2$$

SGD

$$\min_w f(w) := \frac{1}{n}\sum_{i=1}^n f_i(w)$$

i_k sample from $\{1, 2, \cdots, n\}$

$$w^{k+1} = w^k - \alpha_k \nabla f_{i_k}(w^k)$$

Thm. If f is L-smooth, $f(w) \geq f^*$ for any w

$$\mathbb{E}_i[\|\nabla f_i(w)\|^2] \leq G^2$$

Then $$\min\left\{\mathbb{E}[\|\nabla f(w^k)\|^2]\right\}_{k=0}^{k-1} \leq \frac{f(w^0) - f^* + \frac{G^2 L}{2}\sum_{k=0}^{k-1} t_k^2}{\sum_{k=0}^{k-1} t_k}$$

Proof. By L-smooth

$$f(w^{k+1}) \leq f(w^k) \underbrace{- t_k \nabla f(w^k)^T \nabla f_{i_k}(w^k)}_{\text{may not be negative}} + \frac{t_k^2 L}{2}\|\nabla f_{i_k}(w^k)\|_2^2$$

But "on average" (in expectation), we may make progress

$I_{k-1} := (i_1, i_2, \cdots, i_{k-1})$

$$\mathbb{E}_{I_k}[f(w^{k+1})] \leq \mathbb{E}_{I_k}\left[f(w^k) - t_k \nabla f(w^k)^T \nabla f_{i_k}(w^k) + \frac{t_k^2 L}{2}\|\nabla f_{i_k}(w^k)\|_2^2\right]$$

$$= \mathbb{E}_{I_{k-1}}[f(w^k)] - t_k \mathbb{E}_{I_{k-1}}[\|\nabla f(w^k)\|^2] + \frac{t_k^2 L}{2}\underbrace{\mathbb{E}_{i_k}[\|\nabla f_{i_k}(w^k)\|_2^2]}_{\text{2nd-order moment}}$$

$$\leq \mathbb{E}[f(w^k)] - t_k \mathbb{E}[\|\nabla f(w^k)\|^2] + \frac{t_k^2 G^2 L}{2}$$

$$\Rightarrow \sum_{k=0}^{k-1} t_k \mathbb{E}[\|\nabla f(w^k)\|^2] \leq f(w^0) - f^* + \frac{G^2 L}{2}\sum_{k=0}^{k-1} t_k^2$$

$$\Rightarrow \min\left\{\mathbb{E}[\|\nabla f(w^k)\|^2]\right\}_{k=0}^{k-1} \leq \frac{f(w^0) - f^* + \frac{G^2 L}{2}\sum_{k=0}^{k-1} t_k^2}{\sum_{k=0}^{k-1} t_k}$$