Lecture 12: OPTIMIZATION

Ann.
* typed notes Sunday

First-order methods only (only use (Clarke) gradients).

* low per-iteration complexity
  (the high per-iter complexity methods, e.g., Newton, give high accuracy, but that's irrelevant for problems with uncertainty.)
* Seem to have favorable bias (towards predictors which generalize well).

Rough plan

* GD & GF near initialization
* homogeneity ($\sigma(cx) = c\sigma(x)$) & margin maximization & feature learning
* Maybe (?): mean field
* Maybe not:
  * landscape
  * SGD
  * SDE
  * Adam

Precise plan.

Near initialization, we'll consider ("training wheels") (projected) GD

(eventually we'll pick $S = \mathbb{R}^p$ (important))

$w_0 \in S$, thereafter $w_{i+1} := w_i - \eta g_i$ ("gradient"; η: stepsize)

where $g_i = \nabla \hat{R}(w_i)$,

and $\hat{R} = \mathcal{L} \circ H$ ($H: \mathbb{R}^p \to \mathbb{R}^n$ (training set baked in); $\mathcal{L}$: convex loss $\mathbb{R}^n \to \mathbb{R}$)

Examples

① Linear (sanity check): $\mathcal{L}(\hat{z}) = \frac{1}{n}\sum_i \ell(\hat{z}_i) \in \mathbb{R}^n$ ($\ln(1+\exp(-\hat{z}_i))$)

$$H(w) = \begin{bmatrix} -\, y_1 x_1^T \,- \\ | \\ -\, y_n x_n^T \,- \end{bmatrix} w \quad \in \mathbb{R}^{n \times d}$$

$$\partial \hat{R} = \partial H \, \partial \mathcal{L}(H\cdot) \qquad \partial\mathcal{L} = \frac{1}{n}\begin{bmatrix} \ell'(\hat{z}_1) \\ | \\ \ell'(\hat{z}_n) \end{bmatrix}$$

$$\partial H = \begin{bmatrix} | & & | \\ y_1 x_1 & - & y_n x_n \\ | & & | \end{bmatrix} \in \mathbb{R}^{d \times n}$$

② 2-layer network, train inner layer

$\mathcal{L}$ & $\partial\mathcal{L}$ as before

$$H: \mathbb{R}^p \to \mathbb{R}^n; \quad H(w) = \begin{bmatrix} y_1 F(x_1; w) \\ | \\ y_n F(x_n; w) \end{bmatrix}$$

where $F(x; w) = \sum_j a_j \sigma(v_j^T x)$ ($a_j = \pm \frac{1}{\sqrt{m}}$)

$$\partial H(w) = \begin{bmatrix} | \\ \mathrm{vec}\left(\sum_j a_j \sigma'(v_j^T x) x e_j^T\right) \\ | \end{bmatrix} \in \mathbb{R}^{p \times n}$$

($\mathrm{vec}(\cdot) \in \mathbb{R}^p$)

More precise plan

* Gradient methods on $\mathcal{L} \circ H$ near init ($\mathcal{L}$ convex)
  * $\mathcal{L}$ Lipschitz, ∂H "bounded in a certain sense"
* Smooth $\mathcal{L}$ & H
* Smooth $\mathcal{L}$ & H, $\mathcal{L}$ strongly convex ("usual")
* weak implicit bias (remove projections)
* GF
* Perceptron / Polyak-Łojasiewicz

Homogeneity...

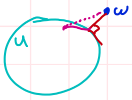
GD $w_{s+1} := w_s - \eta g_s$

"Mirror descent" / "Master" lemma

Let convex closed constraint U be given, & comparator $z \in U$. Then

$$\|w_t - z\|^2 \le \|w_0 - z\|^2 + 2\eta \sum_{s<t} \langle g_s, z - w_s \rangle + \eta^2 \sum_{s<t} \|g_s\|^2,$$

where ↗ is equality if projection never encountered.



"orthogonal projection onto u" $\Pi_U(w) = \arg\min \{ \frac{1}{2}\|w - v\|^2 : v \in U \}$

Proof. For any s

$$\|w_{s+1} - z\|^2 = \|\Pi_U(w_s - \eta g_s) - z\|^2$$

$$\le \|w_s - \eta g_s - z\|^2$$

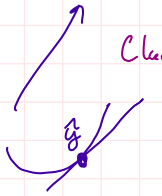
(Π_U is contractive; will show if time)

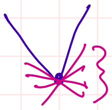
$$= \|w_s - z\|^2 + 2\eta \langle g_s, z - w_s \rangle + \eta^2 \|g_s\|^2,$$

apply $\sum_{s<t}$ to both sides & cancel common term. //

Definition. If $\mathcal{L}$ is convex, then $\forall y, \hat{y}$ "$\mathcal{L}(y) \ge \mathcal{L}(\hat{y}) + \langle \partial \mathcal{L}(\hat{y}), y - \hat{y} \rangle$"

Clarke differential is a set: $\forall y, \hat{y}$, $\forall g \in \partial \mathcal{L}(\hat{y})$, $\mathcal{L}(y) \ge \mathcal{L}(\hat{y}) + \langle g, y - \hat{y} \rangle$.





} all these slopes: $\partial |\cdot| (0)$.

Corollary. U closed convex set, $z \in U$,

$$\varepsilon := \sup_{w \in U} \|H(z) - H(w) - \langle \partial H(w), z - w \rangle\|_\infty,$$

(linearization error from near-init gpe lecture)

$$\sup_{w \in U} \|\partial \mathcal{L}(H(w))\|_1 \le A < \infty,$$

$$\sup_{w \in U} \|\partial H(w)\|_{2,\infty} \le B < \infty, \quad \eta = \frac{\eta_0}{\sqrt{t}}$$

$$\frac{\|w_t - z\|^2}{2\eta_0\sqrt{t}} + \frac{1}{t}\sum_{s<t} \hat{\mathcal{R}}(w_s) \le \hat{\mathcal{R}}(z) + \frac{\|w_0 - z\|^2}{2\eta_0\sqrt{t}} + \frac{A\varepsilon}{\eta_0} + \frac{A^2B^2}{\eta_0\sqrt{t}}.$$

~~$\frac{1}{\sqrt{t}}$~~ $\quad \frac{1}{\sqrt{t}} \quad \varepsilon \approx \frac{1}{\sqrt{t}} \Rightarrow \frac{1}{\sqrt{t}} \quad \frac{1}{\sqrt{t}}$

$\min_{s<t} \hat{\mathcal{R}}(w_s) \le \frac{1}{\sqrt{t}}$

in general can't pick such a z; more natural is $\mathcal{R}(z) \le \inf_{v \in U} \mathcal{R}(v) + \frac{1}{\sqrt{t}}$

$$\Rightarrow \min_{s<t} \hat{\mathcal{R}}(w_s) \le \inf_{v \in U} \hat{\mathcal{R}}(v) + O\left(\frac{1}{\sqrt{t}}\right).$$

Proof. Recall from lemma:

$$\|w_t - z\|^2 \le 2\eta \sum_{s<t} \langle g_s, z - w_s \rangle + \|w_0 - z\|^2 + \eta^2 \sum_{s<t} \|g_s\|^2$$

$$\langle \partial H \partial \mathcal{L}(H(w_s)), z - w_s \rangle = \langle \partial \mathcal{L}(H(w(s))), \partial H^T (z - w_s) \rangle$$

$$= \langle \partial \mathcal{L}(H(w_s)), H(z) - H(w_s) \rangle$$

$$+ \langle \partial \mathcal{L}(H(w_s)), H(w_s) - H(z) + \partial H^T(z - w_s) \rangle$$

$$\le \mathcal{L}(H(z)) - \mathcal{L}(H(w(s))) \text{ by convexity}$$

$$\le \|\partial \mathcal{L}(H(w_s))\|_1 \cdot \varepsilon$$

cannot directly apply convexity

$\|\partial H\|_{2,\infty}$ next time

Example.

② Recall shallow network, $\mathcal{L}$ logistic loss

$$\|\partial \mathcal{L}(H(w))\|_1 = \frac{1}{n}\sum_i |\ell'(H(w)_i)| \le \frac{1}{n}\sum_i 1 = 1 =: A.$$

$$\|\partial H(w)\|_{2,\infty}^2 = \max_i \|y_i \partial F(x_i; w)\|_F^2 = \max_i \sum_j \|a_j \sigma'(v_j^T x_i) x_i\|^2$$

$$= \max_i \sum_j a_j^2 \, 1^2 \, \|x_i\|^2 \le 1.$$

$$\Rightarrow \boxed{\frac{1}{t}\sum_{s<t} \hat{\mathcal{R}}(w_s) \le \hat{\mathcal{R}}(z) + \frac{\|w_0 - z\|^2}{\eta_0\sqrt{t}} + \frac{\varepsilon}{\eta_0} + \frac{1}{\eta_0\sqrt{t}}.}$$

//

Remark. $\frac{1}{m^{1/6}} \approx \varepsilon = \frac{1}{t^{1/2}} \Rightarrow m = t^3$; disaster!

Later proofs we cover get much smaller widths.