DP-GD for Generalized Linear Models

Today

- Generalized Linear Models
- DP-GD, again
- Convergence analysis

Tools for DP optimization

- Better composition
- Convergence analysis
- Privacy amplification by subsampling
- DP-FTRL

GLMs

Prediction task: $(x_1, y_1), \ldots, (x_n, y_n)$

$x_i \in \mathbb{R}^d$

y_i discrete or $\in \mathbb{R}$.

Modeling assumption: our prediction depends on parameters θ through $\langle x_i, \theta \rangle$

loss function: $\ell(\theta, x_i, y_i) = \ell(\langle x_i, \theta \rangle, y_i)$

Example 1 Linear regression, $y_i \in \mathbb{R}$

Predict $\hat{y}_i = \langle x_i, \theta \rangle$

squared loss: $\ell(\theta, x_i, y_i) = (\langle x_i, \theta \rangle - y_i)^2$

Example 2 Logistic regression, $y_i \in \{0, 1\}$

Predict $\hat{y}_i = \dfrac{1}{1 + e^{-\langle x_i, \theta \rangle}}$

cross-entropy loss: $\ell(\theta, x_i, y_i) = -y_i \ln(\hat{y}_i) - (1 - y_i) \ln(1 - \hat{y}_i)$

Claim For some scalar a,

$$\nabla_\theta \ell(\langle x_i, \theta \rangle, y_i) = a x_i$$

Proof Let $z = \langle x_i, \theta \rangle$. For any $j \in [d]$

$$\frac{\partial}{\partial \theta_j} \ell(\langle x_i, \theta \rangle, y_i) = \frac{\partial \ell}{\partial z} \frac{\partial z}{\partial \theta_j}$$

$\frac{\partial \ell}{\partial z}$: index of j

$\frac{\partial z}{\partial \theta_j}$: $\to x_{ij}$

$\boxtimes$

Algorithm 1 (Noisy Gradient Descent)

Input: data $((x_1, y_1), \ldots, (x_n, y_n))$; loss function ℓ; num iterations T; noise scale σ; learning rate η

Output: $\tilde{\theta} \in \mathbb{R}^d$

1. $\theta_0 \leftarrow \vec{0}$
2. For $t = 1, \ldots, T$
3. $\quad \tilde{g}_t \leftarrow \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \ell(\theta_t; x_i, y_i) + Z_t, \quad Z_t \sim \mathcal{N}(0, \sigma^2)$
4. $\quad \theta_{t+1} \leftarrow \theta_t - \eta \tilde{g}_t$
5. Return $\hat{\theta} = \frac{1}{T} \sum_{t=1}^{T} \theta_t$

Privacy?

Claim 2 If $\forall \theta \in \Theta$ and (x, y), $\|\nabla_\theta \ell(\theta; x, y)\|_2 \leq L$, then Alg 1 is (ε, δ)-DP for

$$\sigma = \frac{2L\sqrt{T}\sqrt{\log(1/\delta)}}{n\varepsilon}.$$

How can we be sure gradient is not too big? Clipping!

Def A function $f: \mathbb{R}^d \to \mathbb{R}$ is **L-Lipschitz** if $\forall x, y \quad |f(x) - f(y)| \leq L \cdot \|x - y\|_2$

Fact If $\nabla f(x)$ exists, then $\|\nabla f(x)\|_2 \leq L$.

Accuracy? Need to control effect of noise.

Convergence for GLMs

Theorem (Song, Steinke, Thakkar, Thakurta 2021)

Suppose loss function $\ell(\langle x, \theta \rangle, y)$ is convex in its first parameter and L-Lipschitz over θ $\forall x, y$.

Let $\mathcal{L}(\theta; X, y) = \frac{1}{n} \sum_{i=1}^{n} \ell(\langle \theta, x_i \rangle, y_i)$.

Let $\theta^* = \operatorname{argmin} \mathcal{L}(\theta; D)$ and M the projector to the eigenspace of $\sum_{i=1}^{n} x_i x_i^T$.

For Alg 1 with $\sigma = 2L \frac{\sqrt{T} \sqrt{\log 1/\delta}}{n\varepsilon}$, $T = n^2 \varepsilon^2$,

& some learning rate, we have

$$\mathbb{E}\left[\mathcal{L}(\tilde{\theta}; X, y)\right] - \mathcal{L}(\theta^*; X, y) \leq \frac{L \|\theta^*\|_M \sqrt{1 + 2\operatorname{rank}(M) \cdot \log 1/\delta}}{\varepsilon n}$$

where $\|v\|_M = v^T M v$.

(Walk through this)

$$\eta^* = \sqrt{\frac{\|\theta^*\|_M^2}{T(L^2 + \operatorname{rank}(M)\sigma^2)}}$$

$f(x) = \frac{A}{x} + Bx, \quad x > 0$

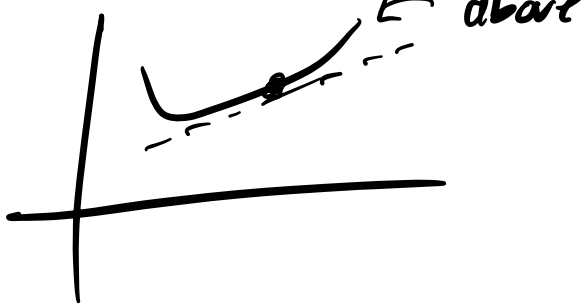
$\Rightarrow x^2 = \frac{A}{B}$

Fact If f is convex & differentiable, then for any x, y,

$$f(x) \geq f(y) + \langle \nabla f(y), u - v \rangle$$

Idea: in 1d, Taylor's thm

$$f(x) \approx f(a) + f'(a)(x-a)$$



Proof $\mathcal{L}(\tilde{\theta}) \overset{\text{Jensen's ineq}}{\leq} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(\theta_t)$

$$\mathcal{L}(\tilde{\theta}) - \mathcal{L}(\theta^*) \leq \frac{1}{T} \sum_{t=1}^{T} \left(\mathcal{L}(\theta_t) - \mathcal{L}(\theta^*) \right)$$

$$\leq \frac{1}{T} \sum_{t=1} \langle \nabla \mathcal{L}(\theta_t), \theta_t - \theta^* \rangle$$

Potential function

$$\Psi_t(\theta) = \mathbb{E}_{z_1, \ldots, z_t} \left[\|\theta - \theta^*\|_M^2 \right]$$

apply to RV.

$$\Psi_t(\theta_{t+1}) = \mathop{\mathbb{E}}_{z_{1:t}}\left[\|\theta_{t+1} - \theta^*\|_M^2\right]$$

$$= \mathop{\mathbb{E}}_{z_{1:t}}\left[\|\theta_t - \eta(g_t + z_t) - \theta^*\|_M^2\right]$$

$$= \mathop{\mathbb{E}}_{z_{1:t}}\left[\|\theta_t - \theta^*\|_M^2 - 2\eta\langle \theta_t - \theta^*, g_t + z_t\rangle + \eta^2\|g_t + z_t\|_M^2\right]$$

$$= \Psi_t(\theta_t) - 2\eta \mathop{\mathbb{E}}_{z_{1:t}}\left[\langle \theta_t - \theta^*, g_t + z_t\rangle\right] + \eta^2 \mathop{\mathbb{E}}_{z_t}\left[\|g_t + z_t\|_M^2\right]$$

(i) (ii) (iii)

i) $\Psi_t(\theta_t) = \Psi_{t-1}(\theta_t)$

ii) $$\mathop{\mathbb{E}}_{z_{1:t}}\left[\langle \theta_t - \theta^*, g_t + z_t\rangle\right] = \mathop{\mathbb{E}}_{z_{1:t-1},\, z_t}\left[\langle \theta_t - \theta^*, g_t\rangle + \langle \theta_t - \theta^*, z_t\rangle\right]$$

$$= \mathop{\mathbb{E}}_{z_{1:t-1}}\left[\langle \theta_t - \theta^*, g_t\rangle\right]$$

iii) $\mathbb{E}_{z_{1:t}}\left[\|g_t + z_t\|_M^2\right] \leq L^2 + \operatorname{rank}(M)\sigma^2$