Lecture 10 Posterior Sampling (Part II) and Objective Perturbation

Yu-Xiang Wang

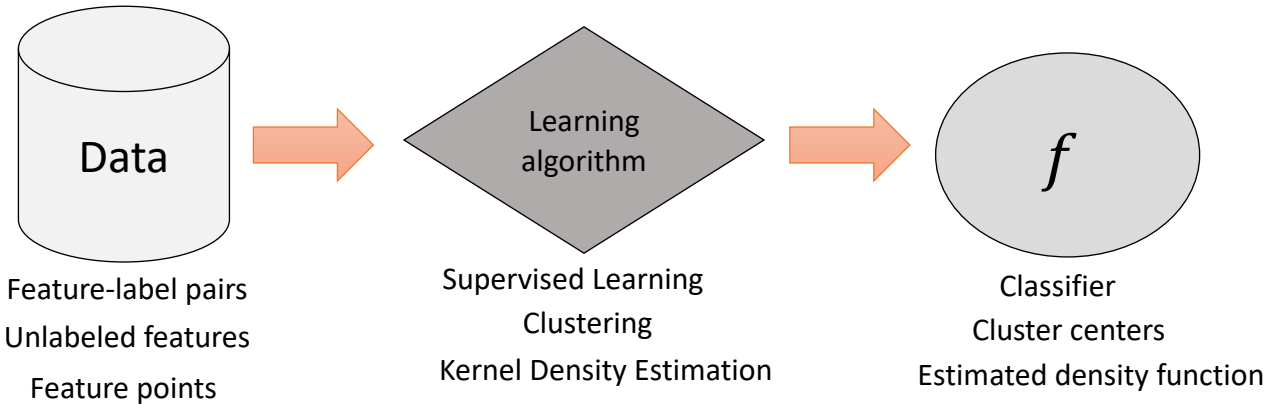
COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

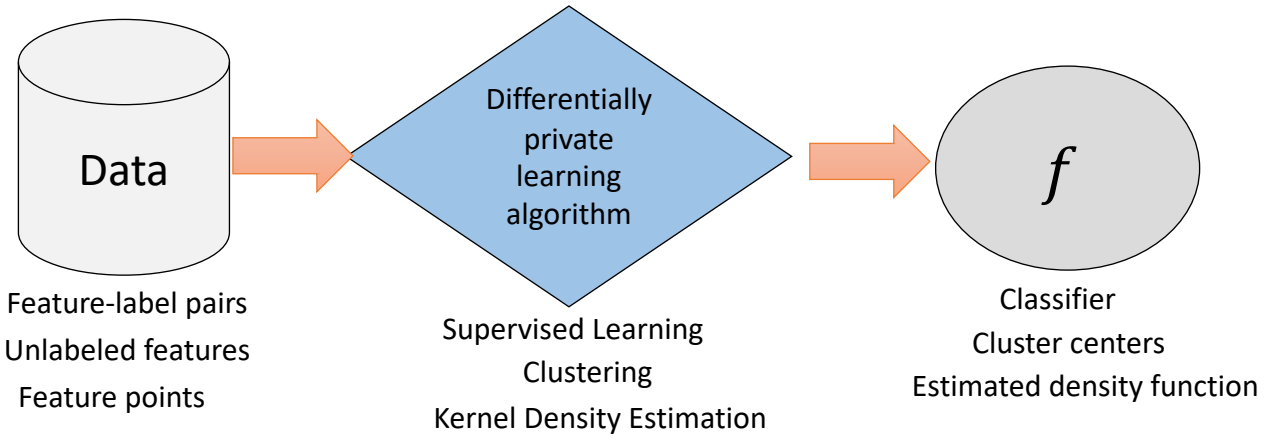
Recap: Last lecture

- Introduction to Differentially Private Machine Learning
  - Problem setup and notations
  - Examples
  - A learning theoretic study of the problem

- Posterior sampling mechanism
  - When the loss functions are bounded
  - A new analysis of the when they are not (not covered)

Recap: Differentially Private Machine Learning





Recap: Statistical Learning Jargons / Notations

- Data space / data points

$$(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$$

- Hypothesis class $\mathcal{H}$
- Loss function, risk and empirical risk

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, (x_i, y_i)) \qquad R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h, (x, y))]$$

- Goal of learning

$$h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(h)$$

Excess Risk

$$R(A(D_n)) - R(h^*) \leq \varepsilon$$

- Realizable vs Agnostic

$R(h^*) = 0$ $\qquad$ $R(h^*) > 0$

Recap: Learning with differential privacy in theory

- Finite hypothesis class

	ERM	Exponential mechanism
Excess risk (Agnostic)	$\sqrt{\frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n}}$	$\sqrt{\frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n}} + \frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n\epsilon}$
Excess risk (Realizable)	$\frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n}$	$\frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n} + \frac{\log(\lvert\mathcal{H}\rvert/\beta)}{n\epsilon}$

- Continuous hypothesis class (bounded VC-dim)
  - No DP algorithm gives consistent learning
  - A Packing lower bound
- However, there are weak assumptions we can add
  - Lipschitz loss functions
  - data-distributions with bounded density

Recap: Packing lower bound in the problem of learning threshold function

- Assume K-data distributions with disjoint support

- Assume a DP algorithm A works on (K-1) of them
  - Expected error < 0.05 => need to stay within support w.p. 0.9

- Implies that it does not work on the remaining one using the definition of DP
  - Probability of outside support > 0.9 => Expected error > 0.45

$$P(A(D_1) \notin [h_1-\tfrac{\eta}{3}, h_1+\tfrac{\eta}{3}]) \geq \sum_{i=2}^{k} P(A(D_1) \in [h_i-\tfrac{\eta}{3}, h_i+\tfrac{\eta}{3}])$$
$$\geq \sum_{i=2}^{k} e^{-n\varepsilon} P(A(D_i) \in [h_i-\tfrac{\eta}{3}, h_i+\tfrac{\eta}{3}])$$
$$\geq \sum_{i=2}^{k} e^{-n\varepsilon} \cdot 0.9 = (k-1)e^{-n\varepsilon} \cdot 0.9$$

Recap: Alternative analysis of the exponential mechanism.

- Bayesian learning:
  - Posterior sampling is private for free, if the log-likelihood function is bounded (or clipped).
  - Utility: $(1+ 4B/\varepsilon)$ multiplicative factor in efficiency.

- More advanced analysis by Minami et al:
  - For strongly log-concave prior

This lecture

- Minami et al's analysis of posterior sampling

- Differentially Private Empirical Risk Minimization
  - Output perturbation
  - Objective perturbation

Optional readings

- “Differentially Private Empirical Risk Minimization” by Chaudhuri, Monteleone, Sarwate.
  - https://www.jmlr.org/papers/volume12/chaudhuri11a/chaudhuri11a.pdf

- “Private Convex Empirical Risk Minimization and High-dimensional regression” by Kifer, Smith and Thakurta
  - http://proceedings.mlr.press/v23/kifer12/kifer12.pdf

Improved analysis of exponential mechanism with strong convexity

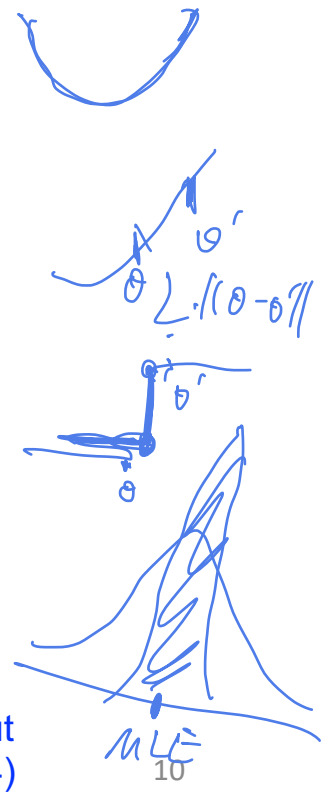
- Assume $\pi(\theta) = e^{-r(\theta)}$ where $r(\theta)$ is μ-strongly convex, i.e., the prior is strongly log-concave.

- Assume that the loss-function is Lipschitz

$$|-\log p(x|\theta) + \log p(x|\theta')| \leq L\|\theta - \theta'\|$$

- Then $\hat{\theta} \sim P(\theta|x_1, ..., x_n) \propto e^{(-\tau \sum_{I=1}^{n} \log p(x_i|\theta) - r(\theta))}$

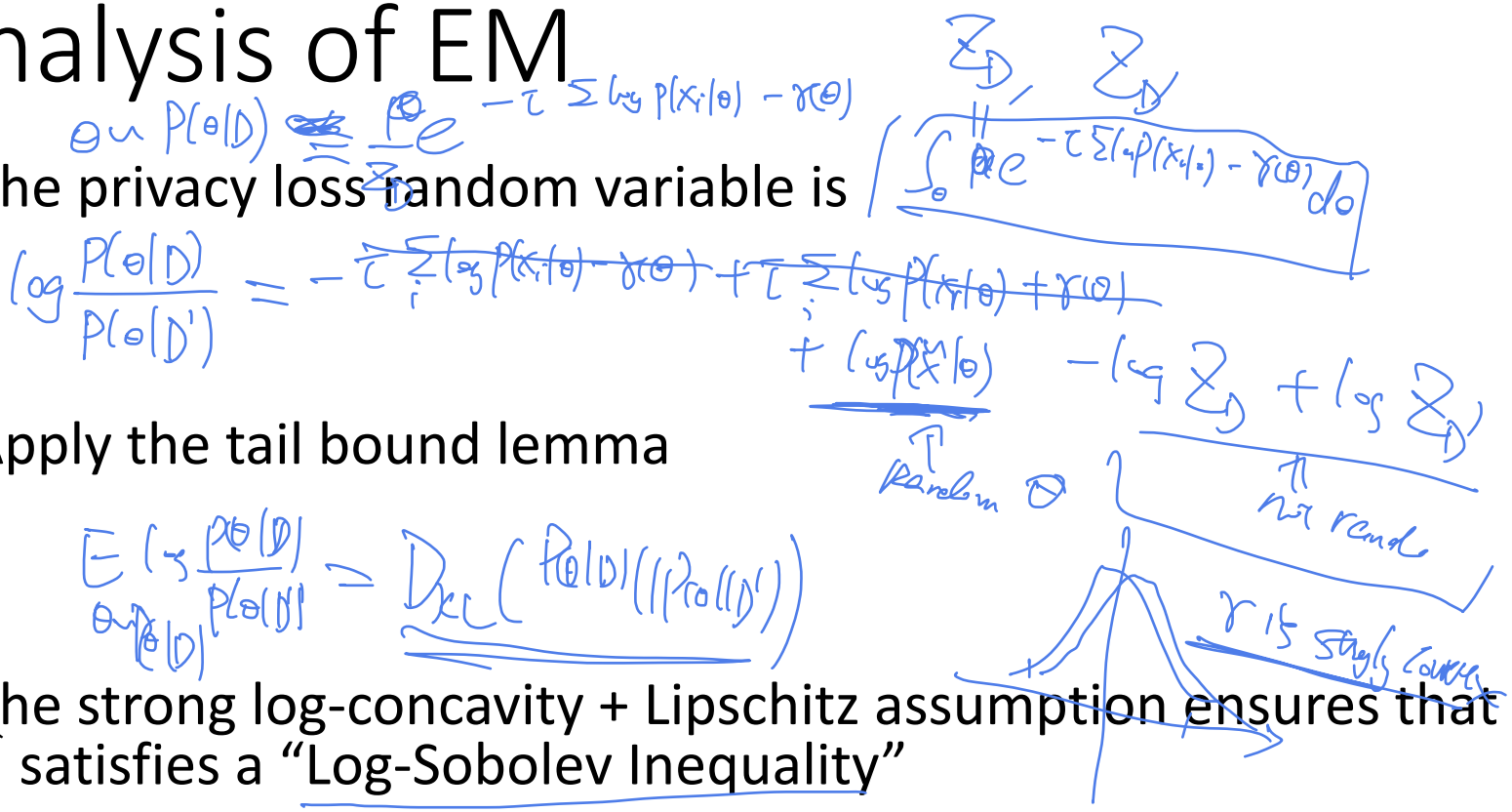
  obeys $(\epsilon, \delta) - \mathrm{DP}$ if $\tau = O(\frac{\epsilon\sqrt{\mu}}{L\sqrt{\log(1/\delta)}})$



Minami, K., Arai, H., Sato, I., & Nakagawa, H. (2016). Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems* (pp. 956-964)

Idea of the proof for the improved analysis of EM

- The privacy loss random variable is



- Apply the tail bound lemma
- The strong log-concavity + Lipschitz assumption ensures that $\hat{\theta}$ satisfies a "Log-Sobolev Inequality"
  - Which ensures a subgaussian-like tail bound for all Lipschitz functions of $\hat{\theta}$
  - And a bound on the KL-divergence.

Minami, K., Arai, H., Sato, I., & Nakagawa, H. (2016). Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems* (pp. 956-964)

The problem of convex empirical risk minimization

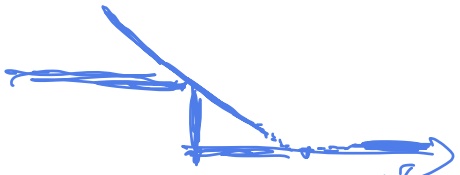
- Data $(x_1, y_1) \cdots (x_n, y_n) \sim D$
  $\in X \times Y = Z$
- Hypothesis class
  $\hat{y} = h(x, \theta) = x^T \theta$
- Loss function
  $\ell(\theta, (x, y))$ ℓ needs to be convex in θ
  $\|$
  $f(x^T \theta, y)$ f is convex $\Rightarrow$ ℓ is convex in θ
  GLM

A specific class of problems of interest: generalized Linear models

- Loss function is of a particular form

$$\ell(\theta, (x, y)) = f(\theta^T x, y)$$

- Examples:
  - Linear regression $(y - x^T\theta)^2$
  - Logistic regression $\log\left(1 + \frac{e^{-yx^T\theta}}{1 + e^{-yx^T\theta}}\right)$ $y = \{-1, 1\}$
  - Support vector machine $y = \{0, 1\}$



Cross Entropy

$$\ell(\theta, (x, y)) = \max\{0, 1 - \theta^T x \cdot y\}$$

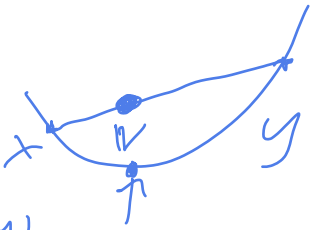
Hinge loss

(Detour) Convex Optimization 101

- Convex functions

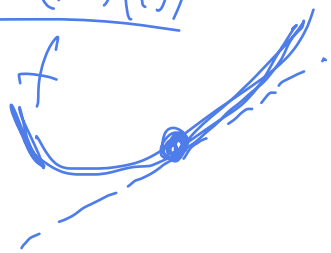
f is convex iff $\forall x, y, 0 \le \alpha \le 1$

$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y)$



For f exists first order conditions: f is convex iff, $\forall x, y$

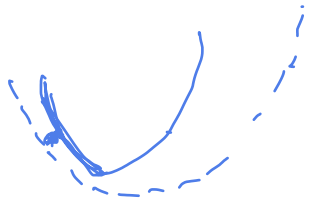
$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle$



- Strongly convex functions

$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\lambda}{2} \|y - x\|^2$

f is λ- Strongly convex



- Optimality condition for differentiable convex functions

$\min_{x \in \mathbb{R}^d} f(x)$ $\quad x^* \in \operatorname{argmin} f(x)$ iff $\nabla f(x^*) = 0$

global — local information

(Detour) Convex Optimization 101

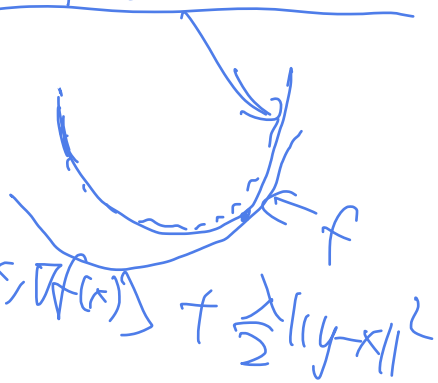
- Lipschitz constant of a function

f is L-Lipschitz if $\forall x, y \quad f(x) - f(y) \leq L \|x - y\|_2$

- Smoothness constant of a function

f is β-smooth if $\forall x, y, \quad f(y) \leq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{\beta}{2} \|y - x\|^2$

λ-Strong convex $f(y) \geq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{\lambda}{2} \|y - x\|^2$



Can we apply Laplace / Gaussian mechanism?

- What is the global sensitivity for the minimizer?

$$\theta^* = \arg\min_\theta \sum_{i=1}^n \ell_i(\theta) \iff \sum_{i=1}^n \nabla \ell_i(\theta^*) = 0$$

$$\tilde{\theta}^* = \arg\min_\theta \sum_{i=1}^n \ell_i(\theta) + \ell(\theta) \qquad \sum_{i=1}^n \nabla \ell_i(\tilde{\theta}^*) + \nabla \ell(\tilde{\theta}^*) = 0$$

Taylor's theorem

$$\sum \nabla \ell_i(\tilde{\theta}^*) = \sum \nabla \ell_i(\theta^*) + \langle \nabla^2 \ell_i(\quad), \tilde{\theta}^* - \theta^* \rangle$$

$$\sum \nabla \ell_i(\tilde{\theta}^*) = -\nabla \ell(\tilde{\theta}^*) \qquad \sum \nabla \ell_i(\theta^*) = 0$$

$$\nabla^2 \ell_i(\quad) \in [\tilde{\theta}^* - \theta^*]$$

$$b = \nabla^2 \ell(\quad)$$

$$Ax = b$$

$$x = A^{-1} b$$

$$\|x\|^2 = b^T A^{-2} b$$

Let's regularize and make it strongly convex

- Regularized objective function

$$\theta_\lambda^* = \arg\min_\theta \sum_{i=1}^n \ell_i(\theta) + \frac{\lambda}{2}\|\theta\|^2 \qquad L(\theta)$$

$$\tilde{\theta}_\lambda^* = \arg\min_\theta \sum_{i=2}^n \ell_i(\theta) + \ell(\theta) + \frac{\lambda}{2}\|\theta\|^2$$

- Now what is the global sensitivity again?

By Strong Convexity

$$L(\tilde{\theta}_\lambda^*) + \frac{\lambda}{2}\|\tilde{\theta}_\lambda^* - \theta_\lambda^*\|^2 \geq L(\theta_\lambda^*) + \frac{\lambda}{2}\|\theta_\lambda^*\|^2 + \frac{\lambda}{2}\|\tilde{\theta}_\lambda^* - \theta_\lambda^*\|^2$$

$$L(\theta_\lambda^*) + \frac{\lambda}{2}\|\theta_\lambda^*\|^2 \geq L(\tilde{\theta}_\lambda^*) + \frac{\lambda}{2}\|\tilde{\theta}_\lambda^*\|^2 + \frac{\lambda}{2}\|\tilde{\theta}_\lambda^* - \theta_\lambda^*\|^2$$

$$+\ell(\theta_\lambda^*) \qquad +\ell(\tilde{\theta}_\lambda^*)$$

add up

$$\lambda\|\tilde{\theta}_\lambda^* - \theta_\lambda^*\|^2 \leq \ell(\theta_\lambda^*) - \ell(\tilde{\theta}_\lambda^*) \leq L\|\tilde{\theta}_\lambda^* - \theta_\lambda^*\|$$

$$\|\tilde{\theta}_\lambda^* - \theta_\lambda^*\| \leq \frac{L}{\lambda}$$

Output perturbation mechanism

1. Solve for the optimal solution

$$\theta^*_\lambda = \arg\min_\theta \; L(\theta) + \frac{\lambda}{2}\|\theta\|^2$$

2. Add noise to the optimal solution

Release $\hat{\theta} = \theta^*_\lambda + N(0, \sigma^2 I_d)$

$\varepsilon < 1$

$$\sigma = \frac{L}{\lambda \varepsilon}\sqrt{2\log\frac{1.25}{\delta}}$$

(ε, δ)-DP

Utility analysis of the output perturbation mechanism

- Let's appeal to the smoothness of the loss function

Objective perturbation

Algorithm 1 Release $\hat{\theta}^P$ via `Obj-Pert` (Kifer et al., 2012)

Input: Dataset D, noise parameter σ, regularization parameter λ, loss function $L(\theta; D) = \sum_i \ell(\theta; z_i)$, convex and twice-differentiable regularizer $r(\theta)$, convex set Θ.
Output: $\hat{\theta}^P$, the minimizer of the perturbed objective.
Draw noise vector $b \sim \mathcal{N}(0, \sigma^2 I)$.
Compute $\hat{\theta}^P$ according to (1).

$$\hat{\theta}^P = \underset{\theta \in \Theta}{\operatorname{argmin}} L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta, \qquad (1)$$

- Remark:
  - Equivalent to adding a random synthetic data point with linear loss
  - One could also get pure DP if we choose b to be:

KKT condition of Objective Perturbation implies a relationship between b and the optimal solution.

$$\hat{\theta}^P = \underset{\theta \in \Theta}{\operatorname{argmin}} L(\theta; D) + r(\theta) + \frac{\lambda}{2}||\theta||_2^2 + b^T\theta,$$

$$\nabla l(\theta; D) + \nabla r(\theta) + \lambda\theta + b = 0$$

$$b(\theta; D) = -\nabla l(\theta; D) - \nabla r(\theta) - \lambda\theta$$

$$b(\theta; D) = -\nabla l(\theta; D) \pm l(\theta) - \nabla r(\theta) - \lambda\theta$$

$$b \sim N(0, \sigma^2 I)$$

$$\log \frac{P(\theta|D)}{P(\theta|D')}$$

Change of variable trick: distribution via the Jacobian

Theorem: Let X, Y be random variables in $\mathbb{R}^d$ and $Y = g(X)$ satisfying that function g is bijective (one-to-one map) and differentiable. Then the probability density function of Y is:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \det \left[\frac{\partial g^{-1}(y)}{\partial y} \right] \right|$$

- Apply this theorem to our problem:

$b \sim N(0, \sigma^2 I_d)$

$b = \theta^{-1}(\theta) = b(\theta; D)$

$$P(A(D) = \hat{\theta}^p) = \frac{1}{(2\pi\sigma^2)^{d}} e^{-\frac{\|b(\hat{\theta}^p; D)\|^2}{2\sigma^2}} \left| \det\left(-\nabla^2 L(\theta; D) - \nabla^2 r - \lambda I_d\right) \right|$$

The privacy loss random variable of the objective perturbation

$$
\begin{aligned}
\log \frac{\Pr\left(\mathcal{A}(D)=\hat{\theta}^P\right)}{\Pr\left(\mathcal{A}(D_{\pm z})=\hat{\theta}^P\right)} &= \log \frac{\left|\det\left(\nabla b(\hat{\theta}^P; D)\right)\right|}{\left|\det\left(\nabla b(\hat{\theta}^P; D_{\pm z})\right)\right|} \frac{\nu(b(\hat{\theta}^P; D); \sigma)}{\nu(b(\hat{\theta}^P; D_{\pm z}); \sigma)} \\
&= \log \frac{\left|\det\left(\nabla b(\hat{\theta}^P; D)\right)\right|}{\left|\det\left(\nabla b(\hat{\theta}^P; D_{\pm z})\right)\right|} + \log \frac{e^{-\frac{1}{2\sigma^2}||b(\hat{\theta}^P; D)||_2^2}}{e^{-\frac{1}{2\sigma^2}||b(\hat{\theta}^P; D_{\pm z})||_2^2}} \\
&= \underbrace{\log \frac{\left|\det\left(\nabla b(\hat{\theta}^P; D)\right)\right|}{\left|\det\left(\nabla b(\hat{\theta}^P; D_{\pm z})\right)\right|}}_{(*)} + \underbrace{\frac{1}{2\sigma^2}\left(||b(\hat{\theta}^P; D_{\pm z})||_2^2 - ||b(\hat{\theta}^P; D)||_2^2\right)}_{(**)}.
\end{aligned}
$$

Bounding (*): the difference of the log determinants

ℓ is β-smooth

$\lambda_i \le \beta$

Lemma 25 (Determinant of Rank-1 perturbation). *For invertible matrix A and vector c, d of compatible dimension*

$$\det(A + cd^T) = \det(A)(1 + d^T A^{-1} c).$$

Recall:

$$b(\hat{\theta}^P; D) = -\left(\nabla \hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla r(\hat{\theta}^P) + \lambda \hat{\theta}^P\right).$$

$$\nabla b(\hat{\theta}^P; D) = -\left(\nabla^2 \hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla^2 r(\hat{\theta}^P) + \lambda I_d\right).$$

- Iterative application of Lemma 25

$$\left|\det\left(\nabla b(\hat{\theta}^P; D_{\pm z})\right)\right| = \left|\det\left(\nabla b(\hat{\theta}^P; D) \mp \nabla^2 \ell(\hat{\theta}^P; z)\right)\right|$$

$$= \left|\det\left(\nabla b(\hat{\theta}^P; D) \mp \sum_{k=1}^{d} \lambda_k u_k u_k^T\right)\right|$$

$$\le \left|\det(\nabla b) \cdot \left(1 + \frac{\beta}{\lambda}\right)^d\right|$$

$$= \left|\det\left(\nabla b(\hat{\theta}^P; D) \mp \lambda_1 u_1 u_1^T \mp \sum_{k=2}^{d} \lambda_k u_k u_k^T\right)\right|$$

$$= \left|\det\left(\nabla b(\hat{\theta}^P; D) + \sum_{k=2}^{d} \lambda_k u_k u_k^T\right) \cdot \left(1 + u_1^T A^{-1} u_1\right)\right| \le \left|\det\left(\nabla b(\hat{\theta}^P; D)\right) \cdot \left(1 + \frac{1}{\lambda}\right)^d\right|$$

$\nabla^2 \ell(\theta; z) = U \Lambda U^T$

- For generalized linear models

GLM $\quad \nabla_\theta \ell(\theta^T x, y) = f'(\theta^T x, y) \cdot x$

$f' \in \mathbb{R}$ $\quad x \in \mathbb{R}^d$

xx^T - scale

$\le \log\left(1 + \frac{\beta}{\lambda}\right)$

Recall:

$$b(\hat{\theta}^P; D) = -\left(\nabla \hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla r(\hat{\theta}^P) + \lambda \hat{\theta}^P\right).$$

$$\nabla b(\hat{\theta}^P; D) = -\left(\nabla^2 \hat{\mathcal{L}}(\hat{\theta}^P; D) + \nabla^2 r(\hat{\theta}^P) + \lambda I_d\right).$$

Bounding (**)

$$(**) = \frac{1}{2\sigma^2}\left(||b(\hat{\theta}^P; D_{\pm z})||_2^2 - ||b(\hat{\theta}^P; D)||_2^2\right)$$

$$= \frac{1}{2\sigma^2}\left[\mp \nabla \ell(\hat{\theta}^P; z)\right]\left[2b(\hat{\theta}^P; D) \mp \nabla \ell(\hat{\theta}^P; z)\right]$$

- For generalized linear models:

Putting everything together

- For general loss functions

Next lecture

- Utility analysis of ObjPert

- Noisy Gradient Descent