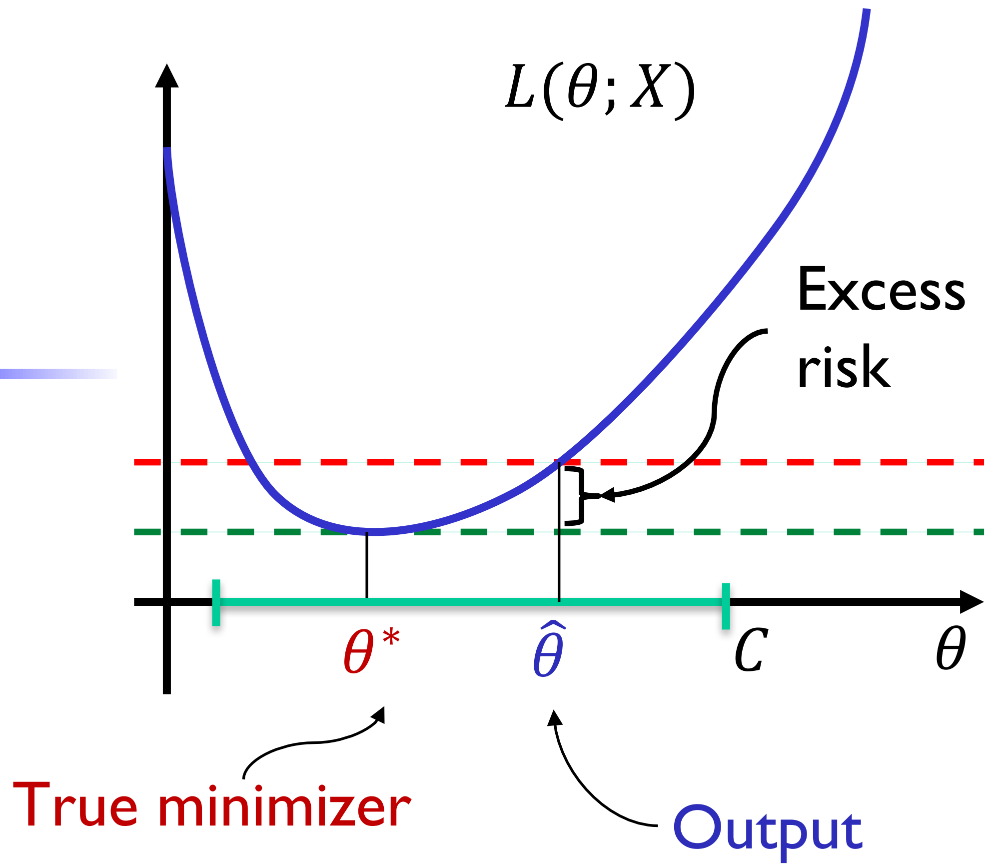
Convex Optimization with Differential Privacy



Adam Smith

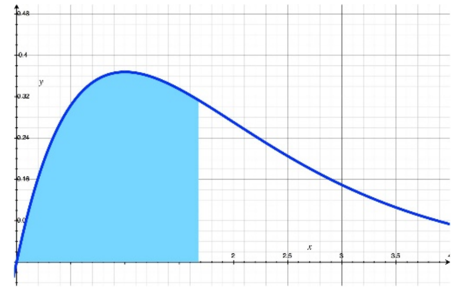
Penn State

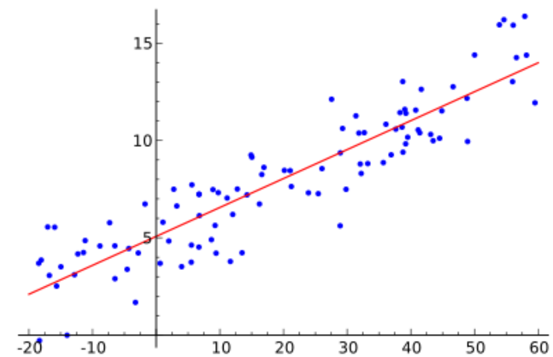
Bar-Ilan Winter School
February 15, 2017

PennState
College of Engineering

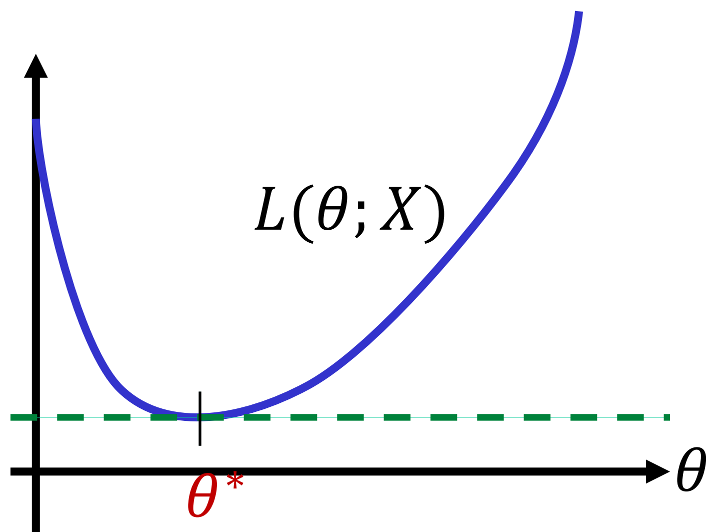
Some Common Computations in Statistics

- Median
- OLS linear regression
- Logistic regression
- Support vector machine





Natural estimator is the result of a minimizing a convex loss function



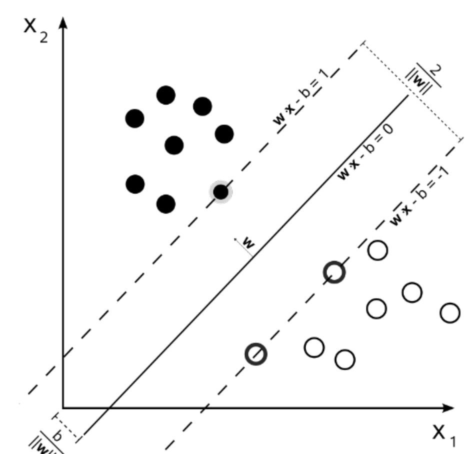


Image: steve-cronin.blogspot.com

Convex Empirical Risk Minimization

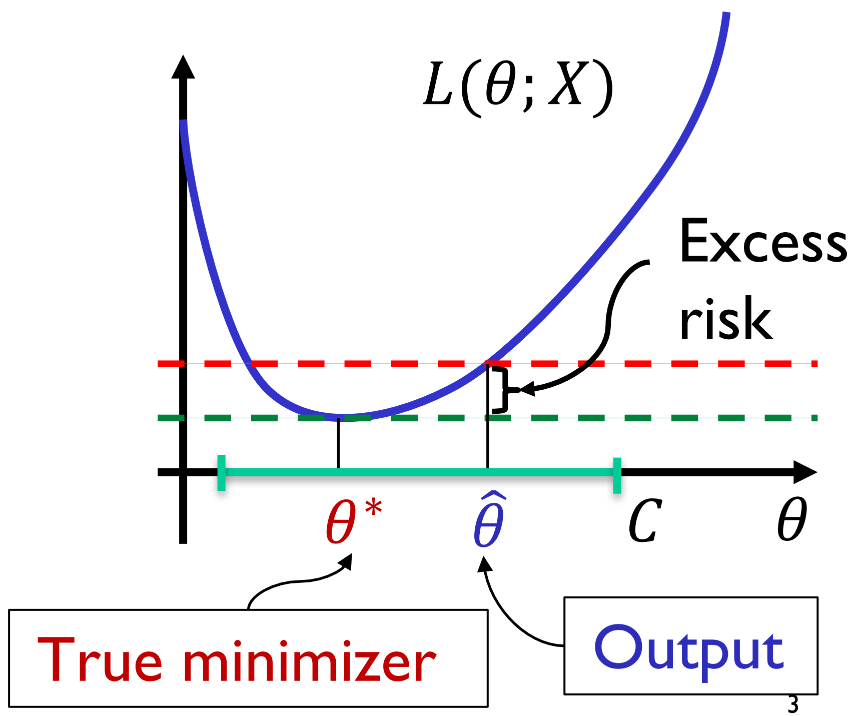
- Data set $X = (x_1, x_2, \dots, x_n) \in U^{\boldsymbol{n}}$
- **Goal:** find a "parameter" $\theta \in C \subseteq \mathbb{R}^{\boldsymbol{d}}$ which minimizes

$$L(\theta; X) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i)$$

where

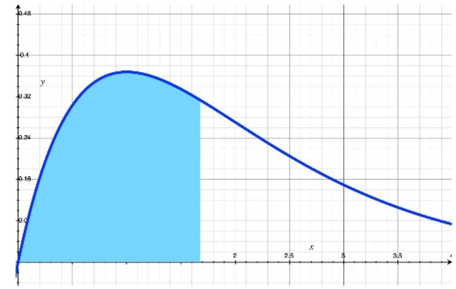
- $\ell(\cdot; x)$ is convex for all x
- C is convex

Goal: small excess risk

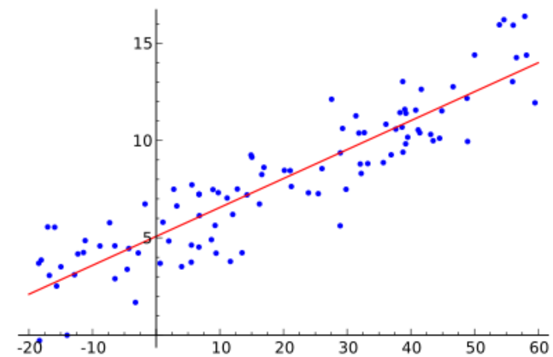


Convex Loss Functions

- Median (in $\mathbb{R}^d$)
  - ➢ $\ell(\theta; x) = \|\theta - x\|_2$
  - ➢ $L(\theta; X) = \frac{1}{n}\sum_i \|\theta - x_i\|_2$



- Linear regression
  - ➢ Data are pairs $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$
  - ➢ $L(\theta; X) = \frac{1}{n}\sum_i (y_i - \langle x_i, \theta\rangle)^2$



- Logistic regression
  - ➢ Data are pairs $(x_i, y_i) \in \mathbb{R}^d \times \{-1,1\}$
  - ➢ $\ell(\theta; x) = \ln(1 + \exp(-y\langle\theta, x\rangle))$
- Support vector machine
  - ➢ $L(\theta; X) = \frac{1}{n}\sum_i h(y_i\langle x_i, \theta\rangle) + \lambda\|\theta\|_2^2$ where $h(z) = (1 - z)_+$

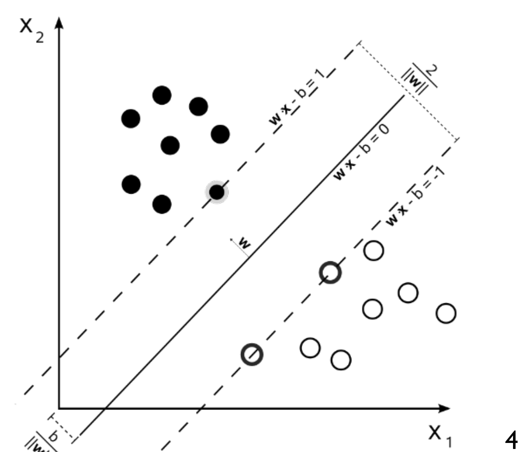
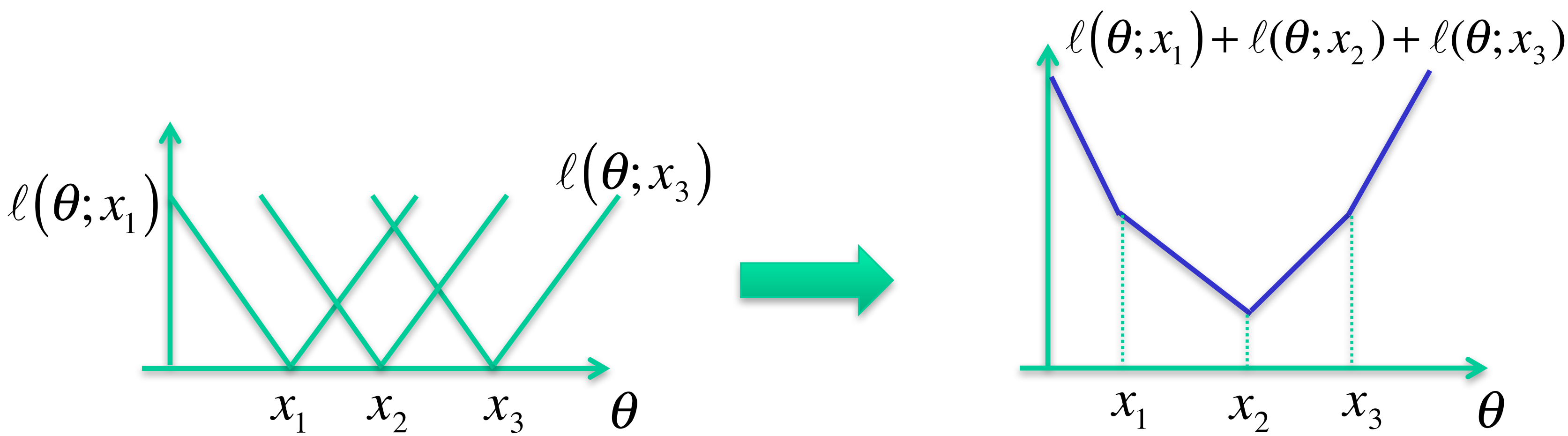


Image: steve-cronin.blogspot.com

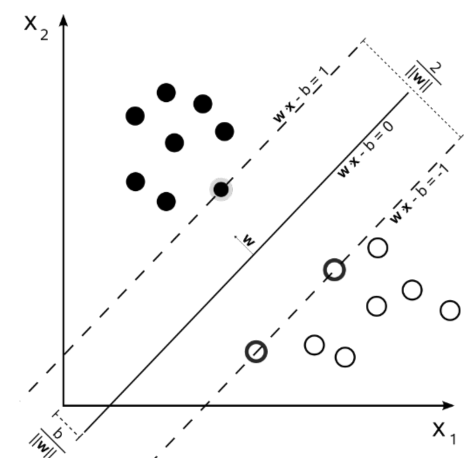
Why care about privacy in ERM?

- Median: Minimizer can be a data point



$\theta^* = x_2$ is the minimizer

- SVM: Dual form of solution encodes high-dimensional data points in the clear



- Reconstruction/membership attacks can use regression parameters

Convex Empirical Risk Minimization

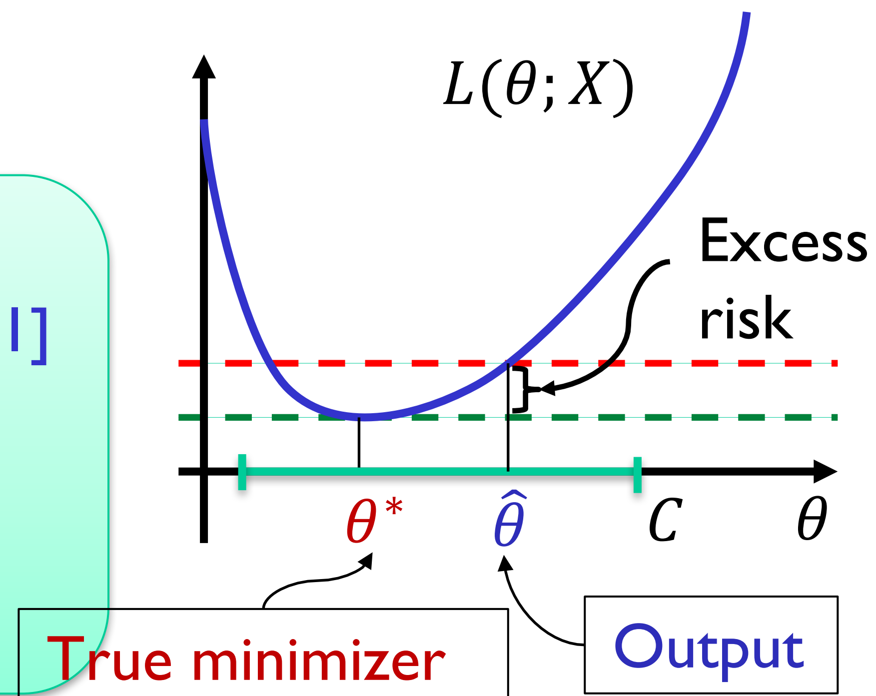
- Data set $X = (x_1, x_2, \ldots, x_n) \in U^{\boldsymbol{n}}$
- **Goal:** find a "parameter" $\theta \in C \subseteq \mathbb{R}^{\boldsymbol{d}}$
  which minimizes $L(\theta; X) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i)$

where

- $\ell(\cdot; x)$ is convex for all x
- C is convex

Goals:

- [Chaudhuri,Monteleoni,Sarwate'11]
  $\hat{\theta} = A(x_1, \ldots, x_n)$ is
  (ϵ, δ)-differentially private
- Small expected excess risk



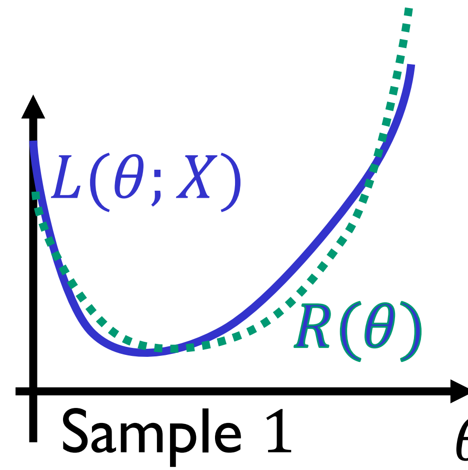
Population versus empirical risk

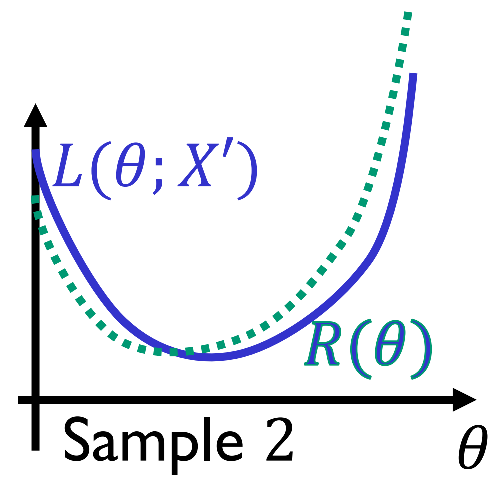
- Suppose $X = (X_1, \dots, X_n) \sim_{i.i.d.} P$
- Empirical risk $L(\theta; X) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i)$
- Population risk (generalization error)

$$R(\theta) = \mathbb{E}_{X\sim_{i.i.d.}P}\big(\mathrm{L}(\theta; X)\big) = \mathbb{E}_{X_i}\big(\ell(\theta; X_i)\big)$$

Goals:

- [Chaudhuri,Monteleoni,Sarwate'11] $\hat{\theta} = A(x_1, \dots, x_n)$ is (ϵ, δ)-differentially private
- Small expected **empirical** risk
- Small expected **population** risk





Population versus empirica

- Suppose $X = (X_1, \dots, X_n) \sim_{i.i.d.} P$
- Empirical risk $L(\theta; X) = \frac{1}{n}\sum_{i=1}^{n} \ell$
- Population risk (generalization er

$$R(\theta) = \mathbb{E}_{X \sim_{i.i.d.} P}\big(\mathrm{L}(\theta; X)\big)$$
$$= \mathbb{E}_{X_i}\big(\ell(\theta; X_i)\big)$$

Bounds depend on assumptions about $\ell, C, R, \dots$

Typical setting:

- ℓ is 1-Lipschitz
- $C \subseteq (\ell_2\ ball)$

Goals:

- [Chaudhuri,Monteleoni,Sarwate'11]
  $\hat{\theta} = A(x_1, \dots, x_n)$ is (ϵ, δ)-differentially private
- Small expected **empirical** risk
- Small expected **population** risk

Difference between these is "generalization error"

Bounded by $\epsilon + \delta$ for DP algorithms!

Focus for now on empirical risk

Lipschitz assumption

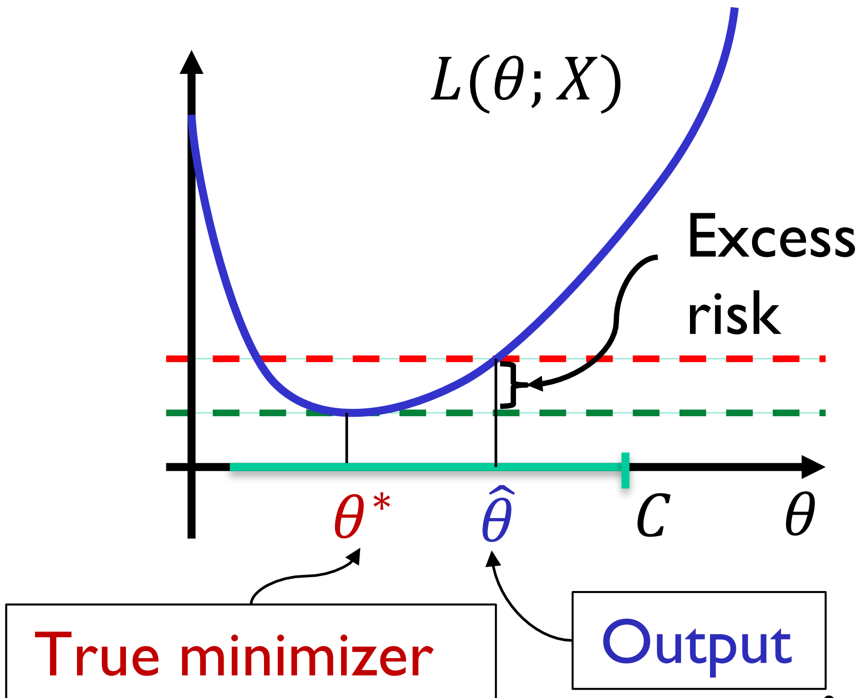
- Data set $X = (x_1, x_2, \ldots., x_n) \in U^{\boldsymbol{n}}$
- **Goal:** find a "parameter" $\theta \in C \subseteq \mathbb{R}^{\boldsymbol{d}}$ which minimizes $L(\theta; X) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i)$

where

- $\ell(\cdot; x)$ is convex for all x
- C is convex

How can we bound each person's influence?

- Assume ℓ is "c-Lipschitz":
  $\|\nabla \ell(\theta; x)\|_2 \leq c$
  for all $x \in U$ and $\theta \in C$.



Bounds *(n individuals, parameter $\theta \in C \subset \mathbb{R}^d$)*

	Privacy	Excess Risk	Technique
Lipschitz	$(\epsilon, 0)$-DP	$O\left(\frac{d}{n} \cdot \frac{1}{\epsilon}\right)$	Exponential sampling [McSherry Talwar, BST14])
	(ϵ, δ)-DP	$O\left(\frac{\sqrt{d}}{n} \cdot \frac{\log\left(\frac{n}{\delta}\right)}{\epsilon}\right)$	Noisy stochastic gradient descent [Williams McSherry, Song Chaudhuri Sarwate, BST14]
Λ-Strongly convex	$(\epsilon, 0)$-DP	$O\left(\frac{d^2}{n^2\Lambda} \cdot \frac{1}{\epsilon}\right)$	Localization (new technique)
	(ϵ, δ)-DP	$O\left(\frac{d}{n^2\Lambda} \cdot \frac{\log^2\left(\frac{n}{\delta}\right)}{\epsilon}\right)$	No

Lower bounds by reduction from attribute proportions [BUV '14]

Assumptions: $||\nabla \ell(\cdot; x)||_2 \leq 1$ for all $x, \theta \in C$

$\text{diameter}(C) \leq 1$

Some Known Algorithms

- Output perturbation [CM'08, RBST'10]
  - Simple but generally suboptimal
- Objective perturbation [CMS'11]
  - Works well for smooth loss functions
- Exponential Sampling [MT'07,BST'14]
  - Works well for $(\epsilon, 0)$-differential privacy
  - Can be tricky to implement
- Noisy SGD [SCS13, BST14, WLF15]
  - Tight bounds for empirical risk minimization
  - Generalizes to mirror descent [TTZ15]
- Noisy Frank-Wolfe [TTZ'15]
  - Exploits structure of constraint set
- Dimension reduction [RBST'10, KJ16,…]
  - Combined with other techniques

Output perturbation [CM'08,RBST'10]

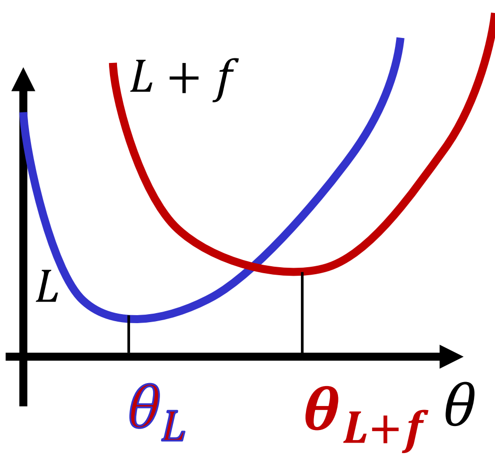
- $L: C \to \mathbb{R}$ is λ-strongly convex if for all $\theta, z \in C$:

$$f(z) \geq f(\theta) + \nabla L(\theta)^T(z-\theta) + \frac{\lambda}{2}\|z-\theta\|_2^2$$

- When $L(\cdot; X)$ is strongly convex, can perturb minimizer $Output = (True\ minimizer) + (noise)$
  - How much noise?
- **Convex Stability Lemma**: If L is λ-strongly convex and f is c-Lipschitz, then

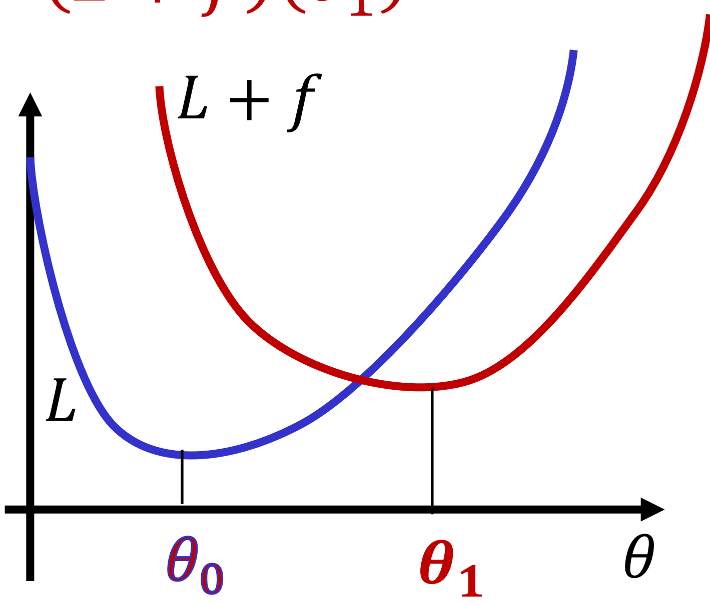
$$\left\|\theta_{L+f}^* - \theta_L^*\right\|_2 \leq \frac{2c}{\lambda}$$

where θ^*'s minimize L and $L+f$ over C.



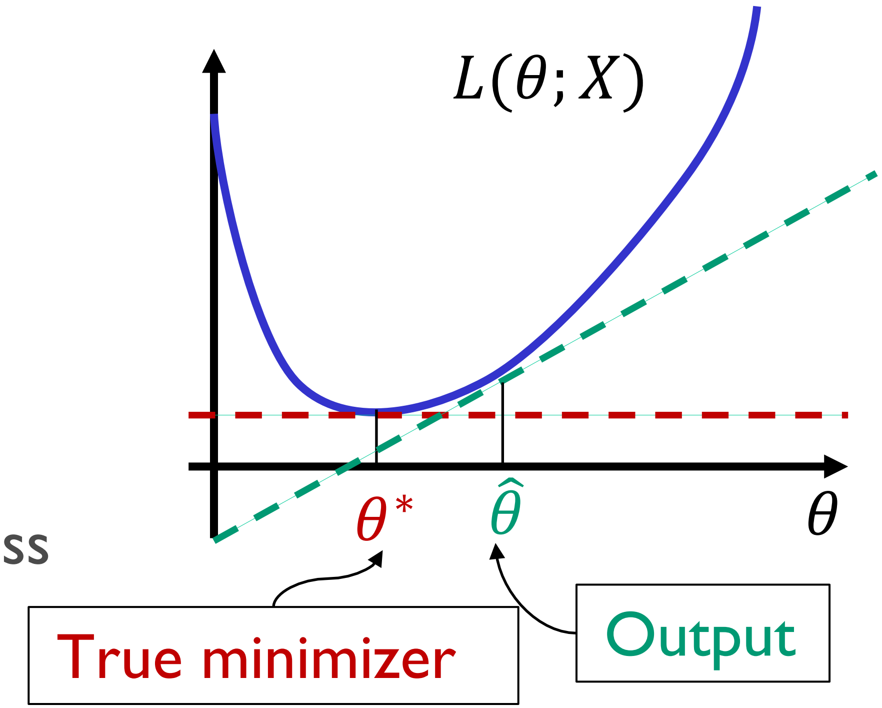
Proof of Strong Convexity Lemma

- $\Delta = \|\theta_1 - \theta_0\|_2$
- By strong convexity: $L(\theta_1) - L(\theta_0) \geq \frac{\lambda}{2}\Delta^2$
- But we also know $L(\theta_1) - L(\theta_0)$
  $\leq L(\theta_1) - L(\theta_0) + (L+f)(\theta_0) - (L+f)(\theta_1)$
  $= f(\theta_0) - f(\theta_1)$
  $\leq c_f \Delta$
- Thus $\frac{\lambda}{2}\Delta^2 \leq c\Delta$
  and so $\Delta \leq 2c/\lambda$.



Objective Perturbation [Chaudhuri,Monteleoni,Sarwate'11]

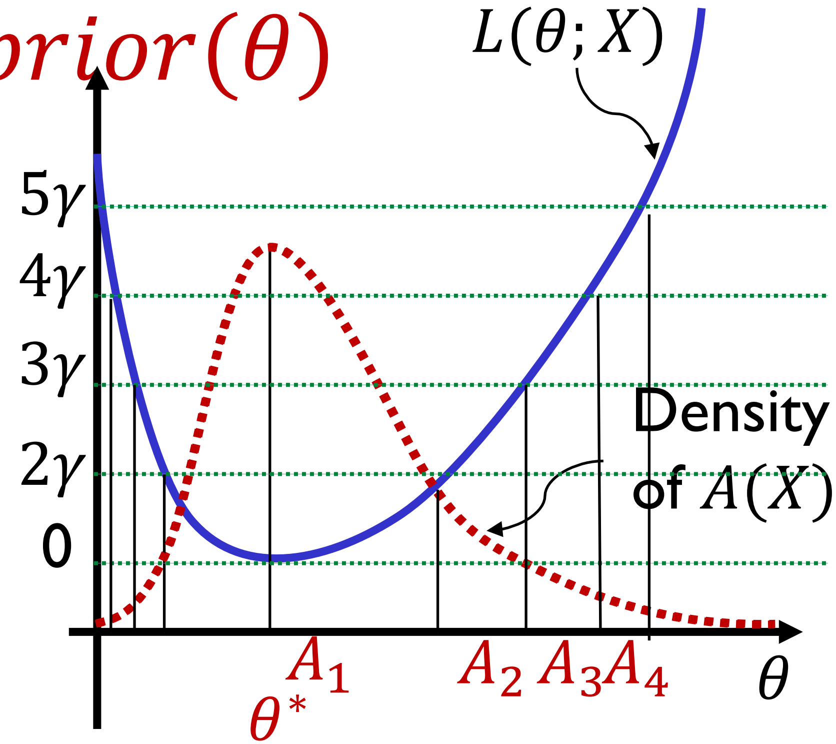
- Pick a random vector $\vec{b} \sim \mathrm{Normal}\left(0, \frac{C}{\epsilon^2 n^2}\right)$
- Return $\hat{\theta} = argmin_{\theta \in C}\, L(\theta; X) + \lambda \|\theta\|_2^2 + \vec{b} \cdot \theta$
  - Where $\lambda \approx 1/\epsilon n$
  - $\nabla L(\hat{\theta}; X) + 2\lambda\hat{\theta} = \vec{b}$
- Privacy requires smooth loss function
  - Median is a counterexample to privacy for nonsmooth loss
  - Nonsmooth regularizer ok (e.g, LASSO)



Exponential Sampling [McSherryTalwar'07, BST'14]

- Define a probability distribution on $\theta \in C$:

$$p(\theta) \propto e^{-\epsilon n \cdot L(\theta;X)} \cdot prior(\theta)$$



- On input X, $A(X)$ outputs a sample $\hat{\theta}$ from p
  - $(\epsilon, 0)$-DP since no single data point has strong effect on $L(\theta;X)$
- Utility analysis via "peeling" argument
  - Use convexity to argue that $p(A_1) \approx 1$ when $\gamma \approx \frac{d}{\epsilon \cdot n}$
  - Polynomial-time algorithm via careful MCMC argument

Noisy stochastic gradient descent

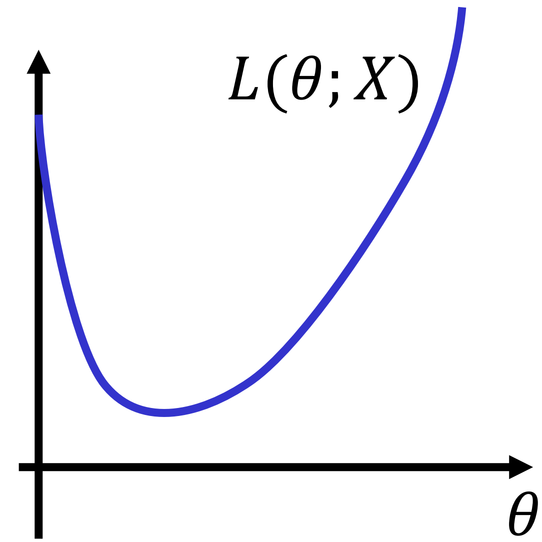
- Run stochastic projected GD, using noisy queries to

$$\nabla L(\theta; X) \approx \frac{1}{n}\sum_i \nabla \ell(\theta; x_i)$$

  - For each step t: pick random i, set

$$g_t = \nabla \ell(\theta_{t-1}; x_i) + (noise)$$

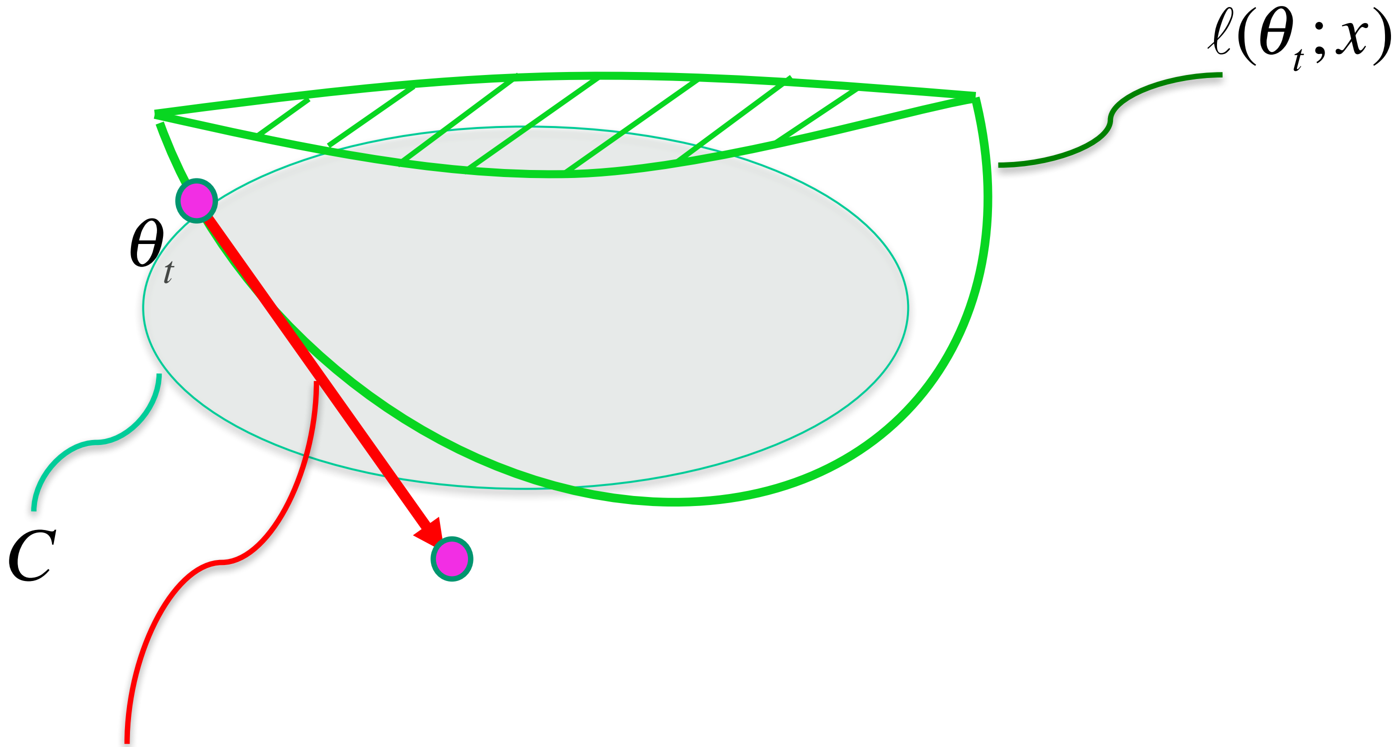
  - Updates: $\theta_t \leftarrow \mathrm{Proj}_C\left(\theta_{t-1} - \frac{1}{\sqrt{t}} g_t\right)$



- Variants evaluated empirically
  [Williams, McSherry '10, Song, Chaudhuri, Sarwate '13]
- Improvements and analysis
  [Bassily, S. , Thakurta'14,
  Abadi, Chu, Goodfellow, McMahan, Mironov, Talwar, Zhang '16]
- Privacy: exploit amplification of ϵ via random sampling
  [Kasiviswanathan, Lee, Nissim, Raskhodnikova, S. '08]

Noisy stochastic gradient descent algorithm

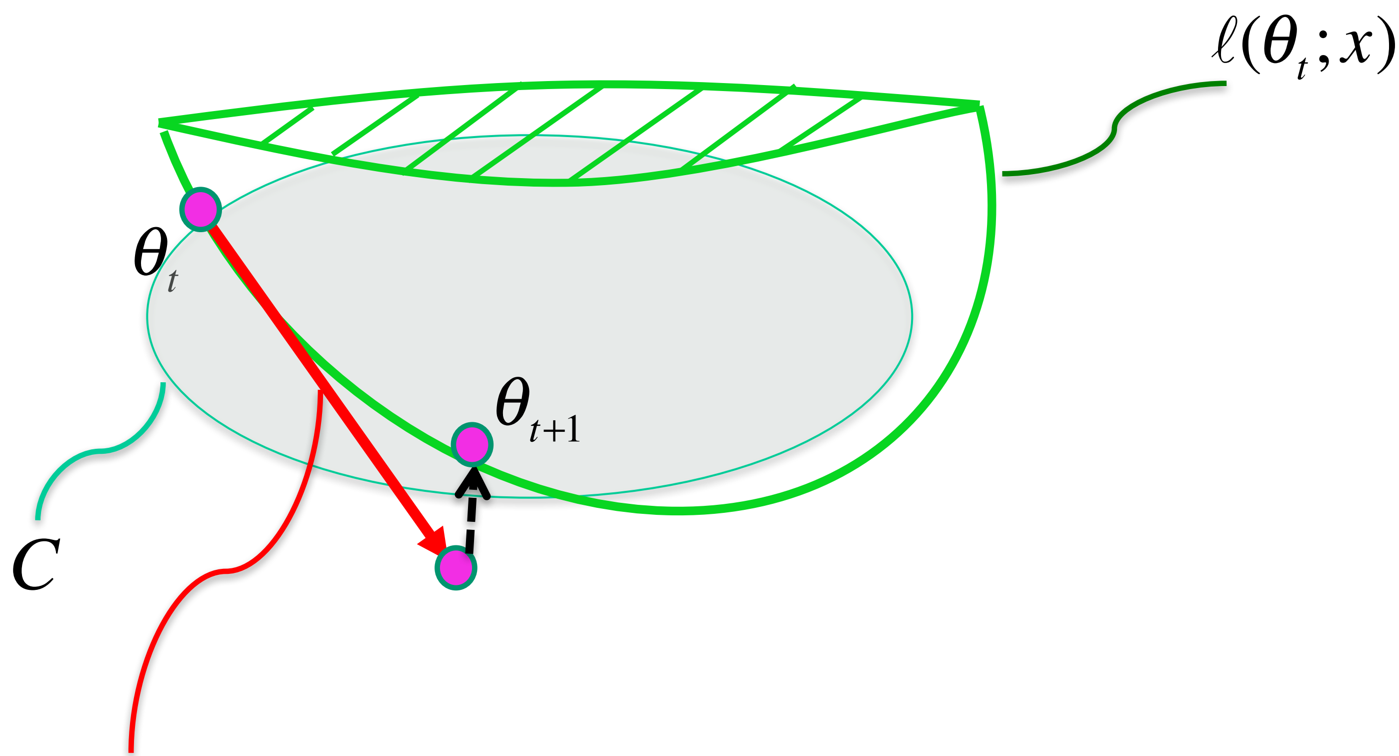
At iteration ***t*** :



direction of the noisy gradient $= -\eta_t \left[n\nabla \ell(\theta_t;x) + b_t \right]$

Noisy stochastic gradient descent algorithm

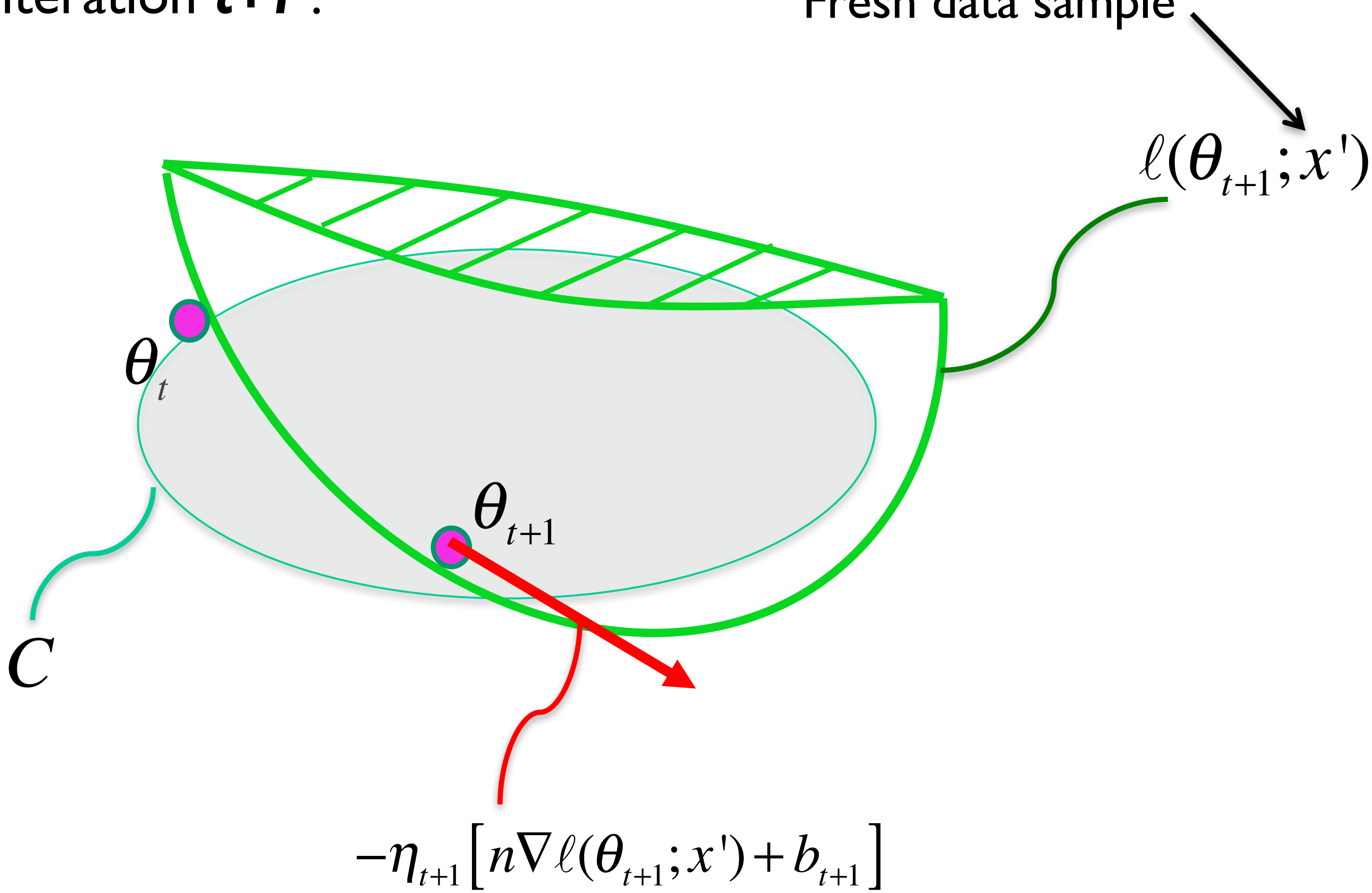
At iteration ***t*** :



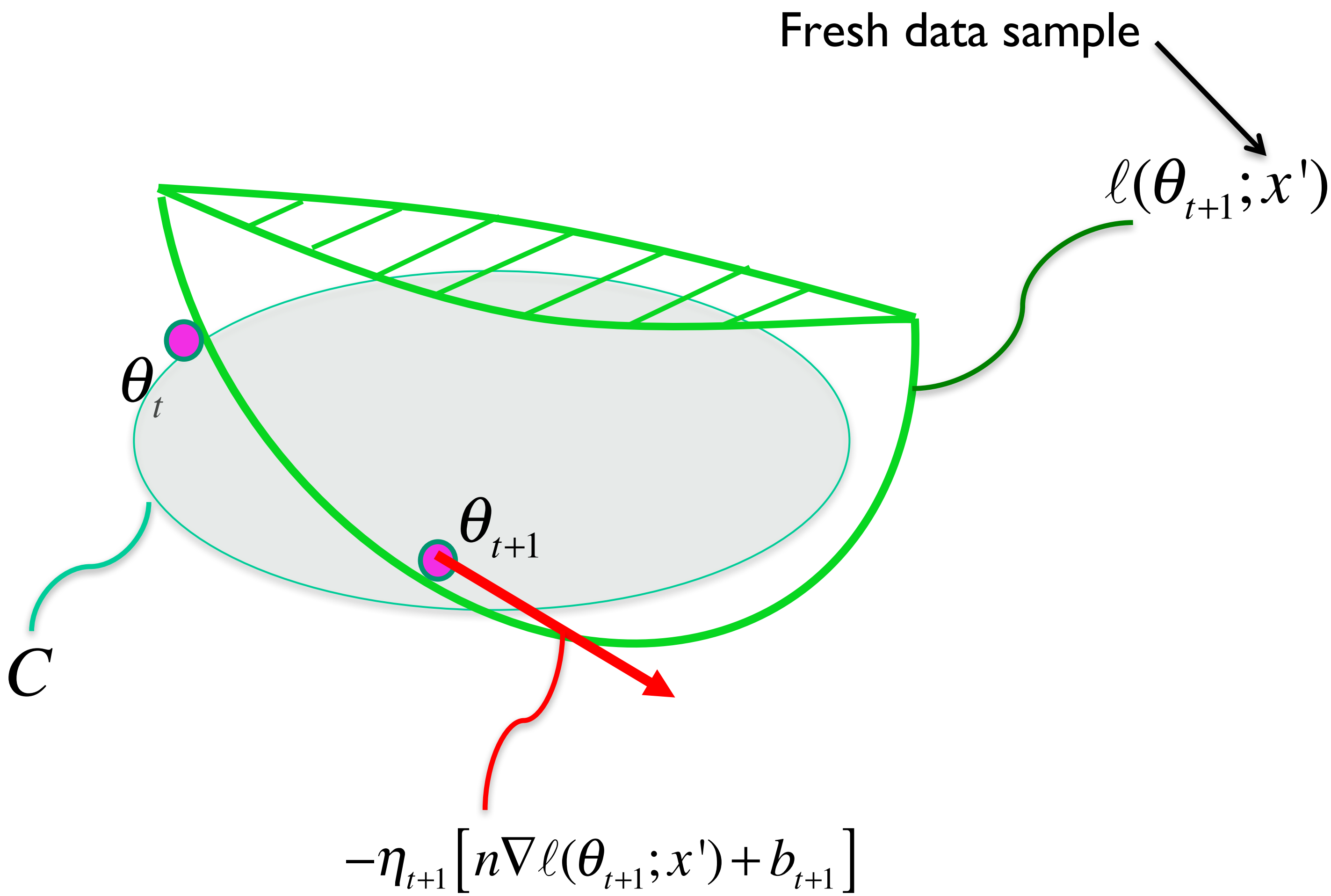
direction of the noisy gradient $= -\eta_t \left[n\nabla \ell(\theta_t; x) + b_t \right]$

Noisy stochastic gradient descent algorithm

At iteration ***t+1*** :



Noisy stochastic gradient descent algorithm



Repeat for n^2 iterations, then output θ_{n^2} .

Noisy Frank-Wolfe *[Talwar, Thakurta, Zhang '15]*

- Suppose C is a polytope (e.g. ℓ_1 ball)
  - $C = ConvexHull(\{vertices\})$
  - (How) can we do optimization over C while leaking little information?

v_{t+1} is always a vertex

- Recall Frank-Wolfe algorithm [FW'56]

$$v_{t+1} = argmax_{v \in C}\left(-v^T \nabla L(\theta_t)\right)$$
$$\theta_{t+1} = \theta_t + \eta(v_{t+1} - \theta_t)$$

- [TTZ15] Use exponential mechanism to select a good vertex
  - If $C \subseteq (\ell_1\ ball)$ and $\|\nabla \ell(\theta; x)\|_\infty \leq 1$, then set

$$v_{t+1} = argmax_{v \in C}\ -v^T\left(\nabla L(\theta_t) + Lap\left(\frac{\sqrt{8\,T \log 1/\delta}}{n\,\epsilon}\right)\right)$$

  - Error grows with $\log(d)$, instead of $poly(d)$

Noisy Frank-Wolfe *[Talwar, Thakurta, Zhang '15]*

- Suppose C is a polytope (e.g. ℓ_1 ball)
  - $C = ConvexHull(\{vertices\})$
  - (How) can we do optimization over C while leaking little information?
- Recall Frank-Wolfe algorithm [FW'56]

v_{t+1} is always a vertex

$$v_{t+1} = argmax_{v \in C}\left(-v^T \nabla L(\theta_t) \right)$$
$$\theta_{t+1} = \theta_t + \eta(v_{t+1} - \theta_t)$$

- [TTZ15] Use exponential mechanism to select a good vertex
  - If $C \subseteq (\ell_1\ ball)$ and $\|\nabla \ell(\theta; x)\|_\infty \leq 1$, then set

$$v_{t+1} = argmax_{v \in C}\ -v^T \left(\nabla L(\theta_t) + Lap\left(\frac{\sqrt{8\,T \log 1/\delta}}{n\,\epsilon} \right) \right)$$

- **For LASSO, get exponential improvement over SGD**

$$\mathbb{E}\left(L(\theta_T) - L(\theta^*)\right) = O\left(\frac{\log(d) + \log(n/\delta)}{(\epsilon n)^{2/3}} \right)$$

The shape of C

- Bounds of [BST'14] apply to any $C \subseteq (\ell_2\ ball)$
  - Assuming ℓ is 1-Lipschitz in $\ell_2$norm
- [TTZ'15] Better bounds when $C \subseteq (\ell_1 ball)$ is a polytope
  - Assuming ℓ has gradients with all entries in $[-1,1]$
- [TTZ'15,KJ'16] Replace $\sqrt{d}$ with Gaussian width

$$w(C) = \mathbb{E}_{b \sim \mathrm{N}(0,\mathbb{I})} \left(\max_{z \in C} |b^T z| \right)$$

  - Exact bounds involve other quantities
- [JT'14] When $C = \mathbb{R}^d$: dimension-independent bounds for regularized GLM's.
- Open problem: Characterize the effect of C.

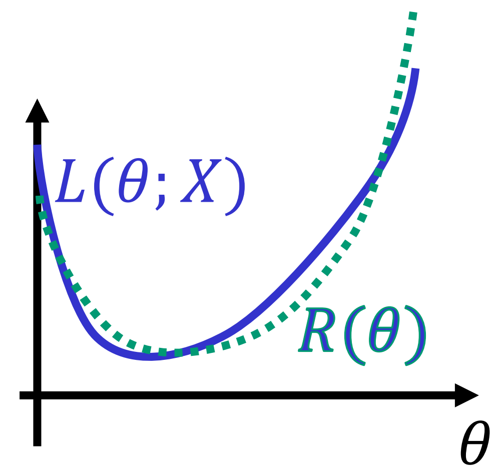
Sequences of Optimization Problems [Ullman15]

- Suppose we have a sequence of optimization problems
  - ℓ_1, C_1
  - ℓ_2, C_2
  - $\ldots$
  - We can solve each one differentially privately
- Naïve accounting shows that privacy loss (ϵ, δ) accumulates as $\sqrt{k}$ for a sequence of k problems
- [U'15, FGV'16] Do better by re-using information
  - View first-order algorithms as sequence of linear measurements
  - Learn model of the data set as you go via private multiplicative weights
  - Privacy loss $poly(\log k, \log|U|, 1/n)$

DP and generalization error

Two methods to control generalization error

- Uniform convergence + ERM
  - Show that with high probability over X, $|L(\theta; X) - R(\theta)|$ small for all θ
  - Not always optimal



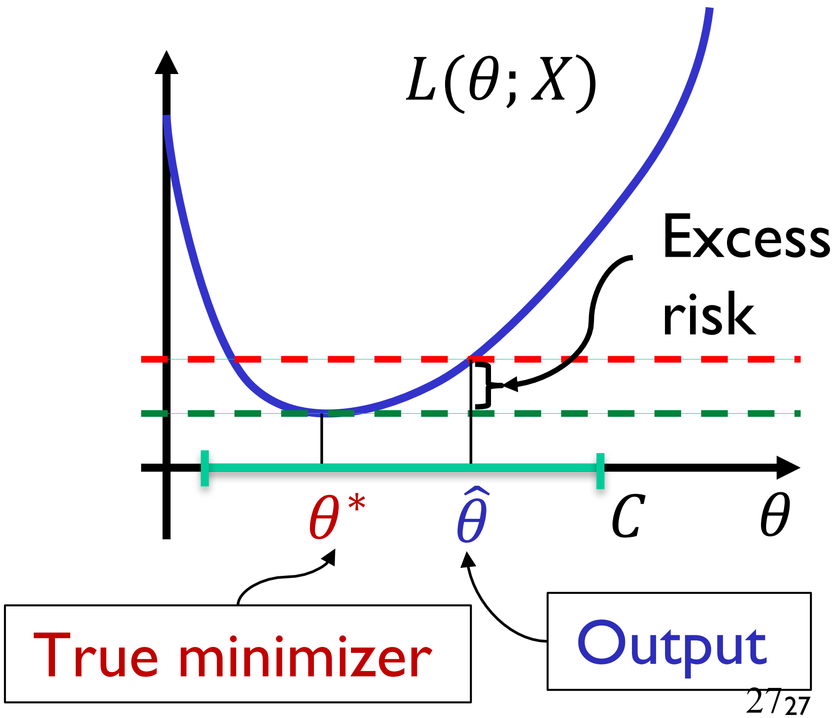
- Stability + ERM
  - Argue that algorithm is “stable”
  - Use general result that for a stable algorithm, $\left|L\left(\hat{\theta}; X\right) - R(\theta)\right|$ is small [Devroye-Wagner 78,…]
  - Every DP algorithm is stable
    - DP algorithms have low generalization error [McSherry’08, BST’14]
    - Used/strengthened in adaptive analysis context by [DFHPRR’15, ...]
  - (gen. error) $\leq$ (empirical error) + $2\epsilon + \delta$ for $\epsilon \leq 1$

Some things I didn't talk about

- Online learning [DNPR'10,JKT'12,ST'13]
- Random projections as a tool for feasibility [KJ'16]
- Approaches for specific loss functions [Refs omitted]
  - Median
  - Linear regression
  - "Robust" statistics
  - Lasso
- Optimization in the local model [KLNRS'08, DJW'13]
- More!

Wrapping up

- State of the art: Algorithms and impossibility results for differentially private convex ERM
  - Tight bounds on (worst-case) empirical risk under general conditions
  - Role of geometry of constraints and functions not fully understood
- Open: Tight bounds on generalization error
- Strong connections between computational efficiency and privacy
- Transfer: Techniques developed for convex optimization also useful for nonconvex objectives [ACGMMTZ'16]



References

- [ACGMMTZ'16] M. Abadi, A. Chu, I. Goodfellow, B. McMahan, I. Mironov, K. Talwar, L. Zhang. Deep Learning with Differential Privacy. *CCS 2016*.
- [BST'14] R. Bassily, A. Smith, A. Thakurta, FOCS 2014.
- [CM'08] K. Chaudhuri, C. Monteleoni. *NIPS 2008*
- [CMS'11] K. Chaudhuri, C. Monteleoni, A. Sarwate, *JMLR* 2011
- [DJW} J. Duchi, M. Jordan, M. Wainwright. FOCS 2013
- [DFHPRR'15] Dwork, Feldman, Hardt, Pitassi Reingold, Roth. STOC 2015.
- [DNPR'10] C. Dwork, M. Naor, T. Pitassi, G. Rothblum. Privacy under continual Observation, *STOC 2010*
- [FGV'16] Feldman, Guzman, Vempala. arXiv:1512.09170
- [JKT'12] P. Jain, P. Kothari, A. Thakurta. Differentially Private Online Learning. *COLT 2012*.
- [JT'14] P. Jain, A. Thakurta. ICML 2014
- [KJ '16] S. Kasiviswanathan, H. Jin. Efficient Private Empirical Risk Minimization for High-dimensional Learning. NIPS 2015
- [KLNRS] S. Kasiviswanathan,
- [KST'13] D. Kifer, A. Smith, A. Thakurta. Private Convex Optimization for Empirical Risk Minimization with Applications to High-dimensional Regression. *COLT 2012*
- [McSherry'08] Frank McSherry. Unpublished. Later appeared as a blog post. https://windowsontheory.org/2014/02/04/differential-privacy-for-measure-concentration/
- [RBHT'10] Rubinstein, Bickell, Huang, Taft, *J. Privacy and Confidentiality* 2011
- [ST'13] A. Smith, A. Thakurta. (Nearly) Optimal Algorithms for Private Online Learning in Full-information and Bandit Settings. *NIPS 2013*
- [Ullman'15] J. Ullman, *PODS 2015*.
- [WM'10] O. Williams, F. McSherry. Probabilistic Inference and Differential Privacy. *NIPS 2010*
- [WLF'15] Wang, Lei, Feinberg. arXiv:1502.07645.