Lecture 11 Noisy Gradient Descent

Yu-Xiang Wang

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

Recap: Last lecture

- Convex empirical risk minimization
- Output perturbation
- Objective perturbation

Recap: Convex ERM and optimality conditions

- Data $(x_1, y_1), ..., (x_n, y_n) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$
- Convex ERM:

$$\min_{\theta \in \Theta \subset \mathbb{R}^d} \sum_{i=1}^{n} \ell(\theta, (x_i, y_i)) =: L(\theta)$$

- Optimality condition: gradient = 0

$$\nabla L(\theta^*) = 0$$

- Assumptions: Lipschitzness, Smoothness

ℓ is L-Lipschitz, β-Smooth

Logistic Regression: 1-Lip, $\frac{1}{4}$-Smooth if $\|x\|_2 \le 1$

Recap: Output perturbation

- Stability of the output via regularization

$$\|\theta^*_\lambda - \theta^*_{\lambda'}\| \le \frac{\|\nabla \ell(\theta^*)\|}{\lambda} \le \frac{L}{\lambda}$$

$\theta^* = \arg\min_\theta L(\theta)$ L

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

- Privacy: from Gaussian mechanism

$$G = \frac{L}{\lambda\varepsilon}\sqrt{2\ln\frac{1.25}{\delta}} \quad (\varepsilon < 1)$$

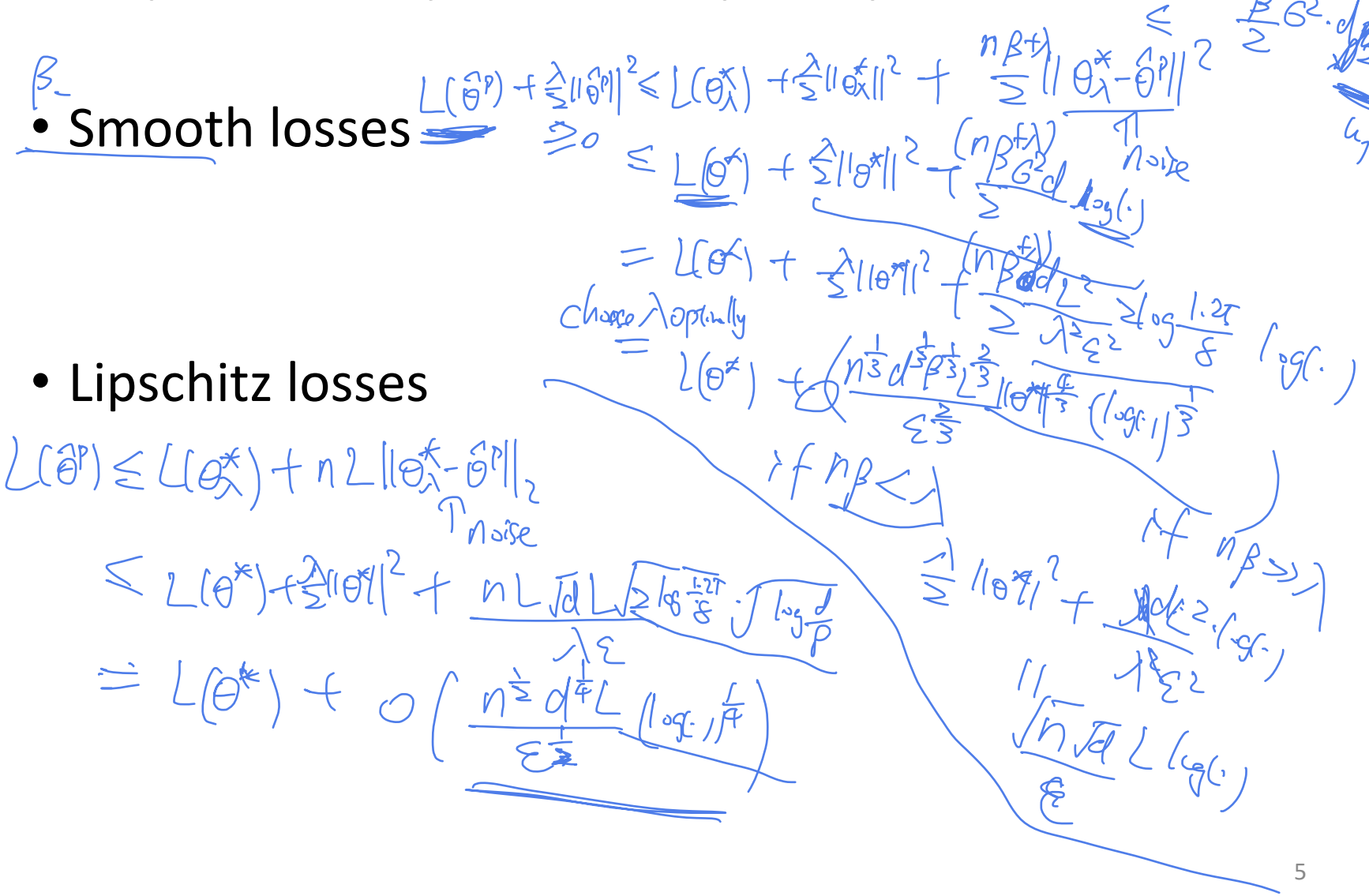
(ε, δ)-DP

- Utility:
  - Last time: under smoothness (has a small error ☹)
  - Let's do it again.

Recap: Utility of Output perturbation

- Smooth losses

- Lipschitz losses



Recap: Objective perturbation

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

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

$\nabla L(\hat{\theta}^P) + \lambda \hat{\theta}^P + b = 0$

- Privacy analysis
  - For GLM

$\log(1+\beta)$ ↓ assume $\lambda > \beta$

$P(PLRV \geq \frac{2\beta}{\lambda} + \frac{L^2}{2\sigma^2} + \frac{\sigma L \sqrt{2\log\frac{2}{\delta}}}{2\sigma^2}) \leq \delta$

$\varepsilon/2$ $\quad$ $\varepsilon/2$

if $\lambda \geq \frac{\varepsilon}{4\cdot\beta}$

we can choose $\sigma = \frac{L}{\varepsilon/2}\sqrt{2\log\frac{1.25}{\delta}}$

  - For General smooth learning problems

$P(PLRV \geq \frac{2d\beta}{\lambda} + \frac{L^2}{2\sigma^2} + \frac{dL\sqrt{2\log\frac{d}{\delta}}}{2\sigma}) \leq \delta$

$\frac{\varepsilon}{2}$ $\quad$ $\frac{\varepsilon}{2}$

$\lambda = \frac{d\varepsilon}{4\beta}$ $\qquad$ $\sigma = \frac{dL}{\varepsilon/2}\sqrt{2\log\frac{1.25d}{\delta}}$

This lecture

- Utility analysis of objective perturbation
- Noisy Gradient Descent
- Privacy amplification by sampling and NoisySGD

Readings

- Chaudhuri et al. / Kifer et al. (continuing)

- Bassily et al. (2014) Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS*. https://arxiv.org/abs/1405.7085
  - For the NoisySGD algorithm
  - For NoisyGD just refer to this lecture note.

Utility analysis of objective perturbation

$\theta_\lambda^* = \arg\min_\theta J(\theta)$

$\theta^* = \arg\min_\theta L(\theta)$

$J(\theta) := L(\theta) + \frac{\lambda}{2}\|\theta\|^2$

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

$\hat{\theta}^p = \arg\min_\theta J(\theta) + b^T\theta$

$J(\hat{\theta}^p) + b^T\hat{\theta}^p \le J(\theta_\lambda^*) + b^T\theta_\lambda^* \quad b^T(\theta_\lambda^* - \hat{\theta}^p)$

$\le J(\theta^*) + b^T(\theta_\lambda^* - \hat{\theta}^p) \leftarrow$

$= L(\theta^*) + \frac{\lambda}{2}\|\theta^*\|^2 + \|b\|_2 \|\theta_\lambda^* - \hat{\theta}^p\|_2 \le L(\theta^*) + \frac{\lambda}{2}\|\theta^*\|^2 + \frac{\|b\|_2^2}{\lambda}$

Recall J is λ-strongly convex

$\|\theta_\lambda^* - \hat{\theta}^p\|_2 \le \frac{\text{Lipschitz}(b^T\theta)}{\lambda} = \frac{\|b\|_2}{\lambda}$

$\le L(\theta^*) + \frac{\lambda}{2}\|\theta^*\|^2 + \frac{\sigma^2 \log\frac{d}{p}}{\lambda}$

choose λ

$= L(\theta^*) + \epsilon \cdot (\text{e.g. } \theta) \frac{1}{2}\|\theta^*\|$

For GLM: $O\left(\frac{\sqrt{d}\, L}{\epsilon}\sqrt{2\log\frac{1}{\delta}}\,\|\theta^*\|\sqrt{\log\frac{d}{p}}\right)$

for general smooth $O\left(\frac{dL}{\epsilon}\sqrt{2\log\frac{1}{\delta}}\,\|\theta^*\|\cdots\right)$

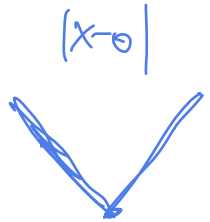
Checkpoint: Compare the excess empirical risk of Output/Objective Perturbation

	Lipschitz losses	Smooth losses	Smooth / Lipschitz GLM
Output Pert	$\frac{d^{1/4} L \|\theta^*\| \log(\frac{1}{\delta})^{1/4}}{n^{1/2}\epsilon^{1/2}}$	$\frac{d^{1/3}\beta^{1/3} L^{2/3} \|\theta^*\|^{4/3} \log(\frac{1}{\delta})^{1/3}}{n^{2/3}\epsilon^{2/3}}$	Same as left
ObjPert	Not applicable	$\frac{dL\|\theta^*\|\sqrt{\log(\frac{1}{\delta})}}{n\epsilon}$	$\frac{\sqrt{d}L\|\theta^*\|\sqrt{\log(\frac{1}{\delta})}}{n\epsilon}$

Lower order terms and dependence on β hidden.

- Normalized by 1/n to be consistent with prior tables.
- Non-private excess risk is on the order of $\sqrt{d/n}$
- Could be $O(d/n)$

What are not quite satisfactory?



- Require the loss to be twice differentiable
  - Convex losses need not be even differentiable

- We did not handle the constrained convex ERM

$$\min \sum_i \ell_i(\theta) \quad \text{s.t. } \|\theta\|_2 \le B$$

- They do not handle non-convex ERM problems, e.g., those that arise when optimizing deep neural networks

Gradient Descent

- Unconstrained, differentiable optimization problem

$$\min_x \; f(x)$$

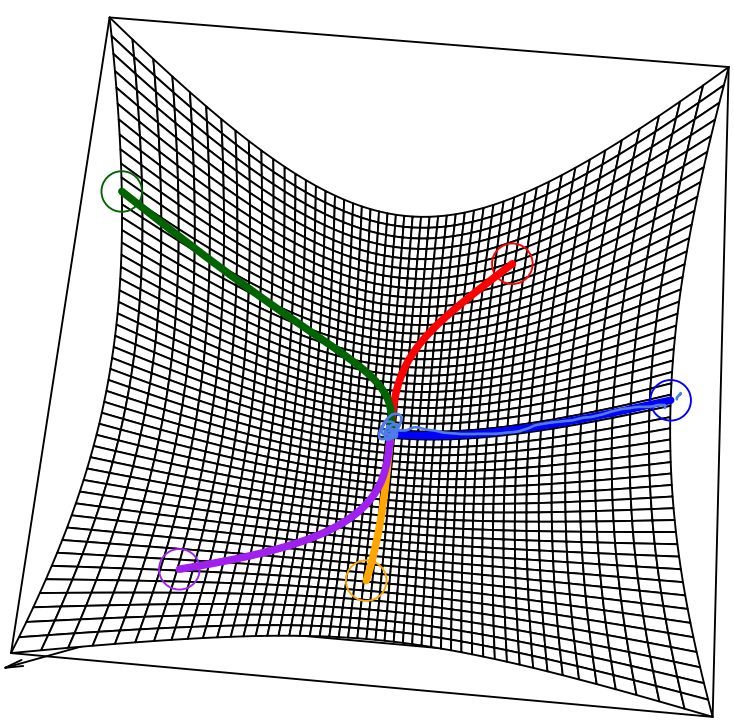
- The algorithm:

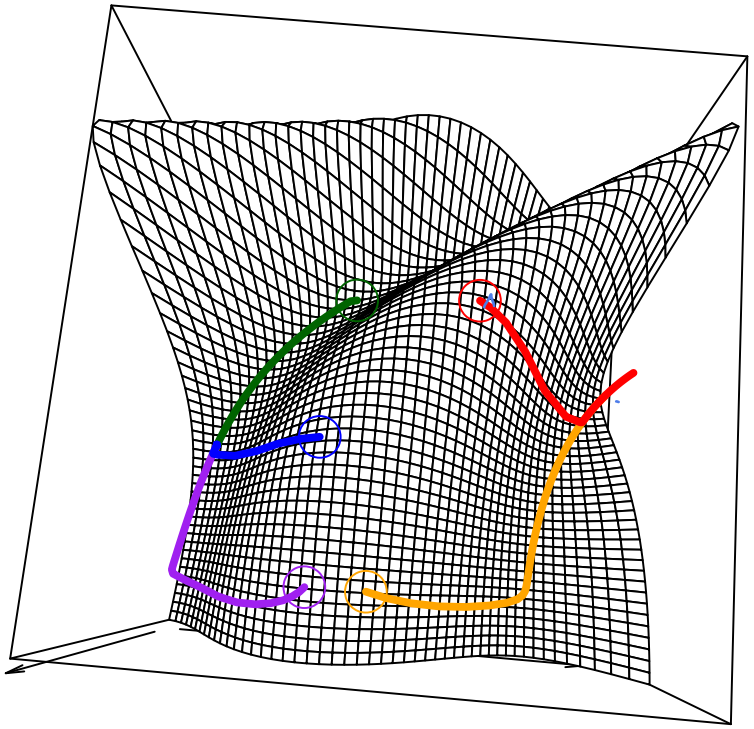
Gradient descent: choose initial point $x^{(0)} \in \mathbb{R}^n$, repeat:

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

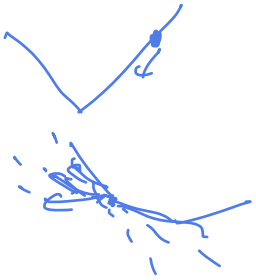
Stop at some point

Gradient descent in convex problems vs nonconvex problems





Extensions of Gradient Descent



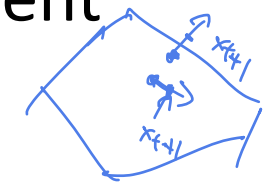
- Non-differentiable case: Subgradient descent

$$\theta_{t+1} = \theta_t - \eta_t g_t$$

(g_t: subgradient)

- Constrained case: Projected gradient descent

$$\theta_{t+1} = P_{\Theta}(\theta_t - \eta_t g_t)$$



- Non-smooth penalty function: Proximal gradient descent

$$\min \sum_{i=1}^{n} l_i(\theta) + r(\theta) + \frac{\lambda}{2}\|\theta\|_2^2$$

($r(\theta)$: non-smooth)

$$\theta_{t+1} = \text{Prox}_r(\theta_t - \eta_t g_t)$$

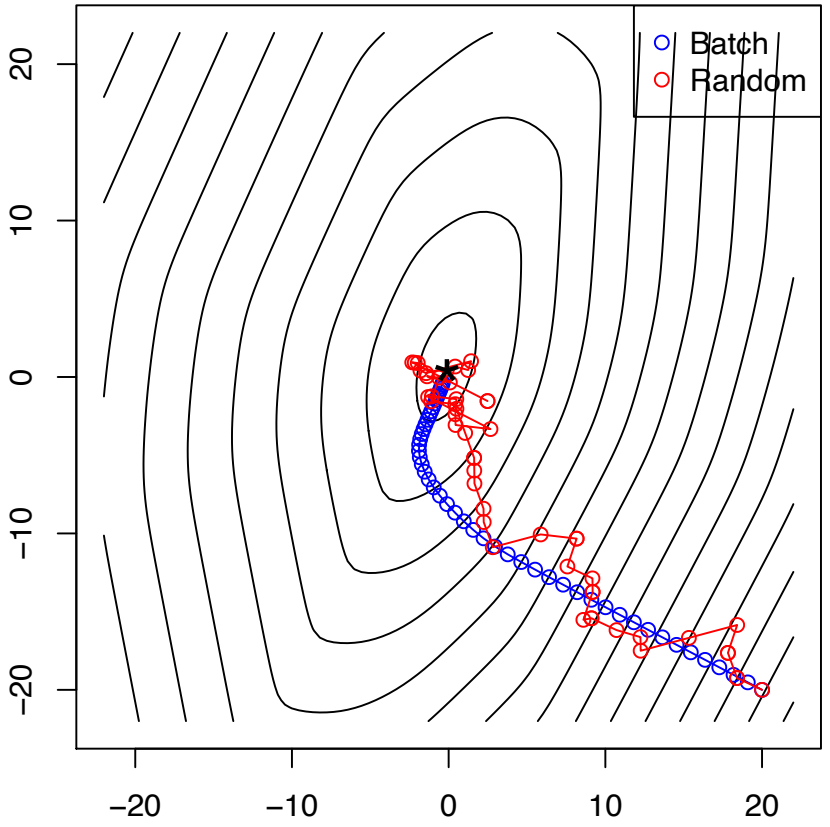
- Nonconvex cases: We give up theoretical guarantees but in practice it works (remarkably well)

Stochastic gradient descent

- Update rule:

$$\theta_{t+1} = \theta_t - \eta_t \hat{g}_t$$

- Assumptions:



The convergence of GD and SGD

- GD in Smooth / convex problems
- GD in general convex problems
- SGD in general convex problems
- SGD in strongly convex problems
- Projected version

Convergence of stochastic gradient descent (in the smooth / nonconvex case)

- Descent Lemma

$\rightarrow x_{t+1} = x_t - \eta_t \hat{g}_t$

① $\mathbb{E}[\hat{g}_t | \theta_{x_t}] = \nabla f(x_t)$

② $\mathbb{E}[\|\hat{g}_t - \mathbb{E}[\hat{g}_t]\|^2 | \theta_{x_t}] \le \sigma^2$

$$f(x_{t+1}) \le f(x_t) + \langle x_{t+1} - x_t, \nabla f(x_t) \rangle + \frac{\beta}{2} \|x_{t+1} - x_t\|^2$$

$$= f(x_t) - \eta_t \langle \hat{g}_t, \nabla f(x_t) \rangle + \frac{\beta}{2} \eta_t^2 \|\hat{g}_t\|^2$$

$$\mathbb{E}[f(x_{t+1}) | x_t] \le f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \frac{\eta_t^2 \beta}{2} (\|\nabla f(x_t)\|^2 + d\sigma^2)$$

$V_{\eta}(x) = \mathbb{E}_x^2 f(x)^2$

$\eta_t = \eta \le \frac{1}{\beta}$

$$= f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{\eta^2 \beta \sigma^2 d}{2}$$

$$= f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|^2 + \frac{\eta^2 \beta \sigma^2 d}{2}$$

$$\mathbb{E}[f(x_{t+1})] \le \mathbb{E}[f(x_t)] - \frac{\eta}{2} \mathbb{E} \|\nabla f(x_t)\|^2 + \frac{\eta^2 \beta^2 \sigma^2 d}{2}$$

add up $t = 1, \dots, T$

Convergence of stochastic gradient descent (in the smooth / nonconvex case)

- Descent Lemma

$$\mathbb{E} f(x_{t+1}) \le \mathbb{E} f(x_t) - \frac{\eta}{2}\mathbb{E}[\|\nabla f(x_t)\|^2] + \frac{\beta\eta^2}{2}d\sigma^2$$

add up

$$\mathbb{E} f(x_{t+2}) \le \mathbb{E} f(x_{t+1}) - \frac{\eta}{2}\mathbb{E}[\|\nabla f(x_{t+1})\|^2] + \frac{\beta\eta^2}{2}d\sigma^2$$

$$\vdots$$

$$\mathbb{E} f(x_T) \le \mathbb{E} f(x_{T-1}) - \cdots$$

$$\Rightarrow \frac{\eta}{2}\mathbb{E}\left(\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\right) \le \mathbb{E} f(x_1) - \mathbb{E} f(x_T) + \frac{T\beta\eta^2}{2}d\sigma^2$$

$$\Rightarrow \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x_t)\|^2\right] \le \frac{2(f(x_1)-f(x^*))}{T\eta} + \beta\eta d\sigma^2 = \frac{2(f(x_1)-f(x^*))\beta\sigma^2}{\sqrt{T}}$$

$$\min_t \|\nabla f(x_t)\|^2 \le$$

Noisy Gradient Descent Mechanism for Convex ERM

- The algorithm:

$$\theta_{t+1} = \theta_t + \eta_t \left[\sum_{i=1}^{n} \nabla l_i(\theta_t) + N(0, G^2 I_d) \right]$$

$\nabla f(\theta_t)$ $\quad$ g_t

- Privacy analysis:
  - A composition of T Gaussian mechanisms

GS: $\nabla f(\theta_t)$ is L $\quad$ l_i is L-Lipschitz

ρ-zCDP: $\quad \rho = \frac{L^2}{2G^2}$

T-gaussian Mech $\quad T\rho$-zCDP

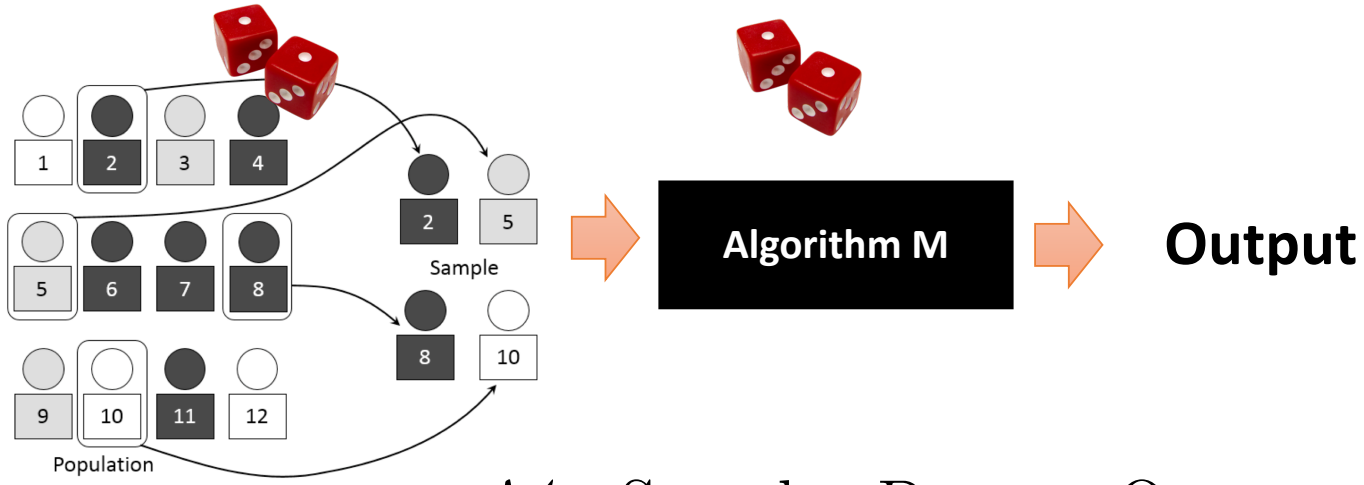
$$\rho_{tot} = \frac{TL^2}{2G^2}$$

$$\frac{G^2}{T} = \frac{L^2}{2\rho_{tot}}$$

$$\mathbb{E}[g_t | \theta_t] = \nabla f(\theta_t)$$

$$\mathbb{E}[\|g_t - \nabla f(\theta_t)\|^2 | \theta_t] = d G^2$$

Privacy Amplification by Sampling



$$\mathcal{M} \circ \text{Sample} : \text{Data} \rightarrow \text{Output}$$

Subsampling Lemma: If M obeys (Ɛ,δ)-DP, then M ∘ Subsample obeys that (Ɛ',δ')-DP with $\delta' = \gamma\delta$

$$\epsilon' = \log(1 + \gamma(e^{\epsilon} - 1)) = O(\gamma\epsilon)$$

Random subset sampling vs Poisson sampling

The Noisy Stochastic Gradient Descent Mechanism (NoisySGD)

- Privacy analysis:
  - A composition of T subsampled gaussian mechanism.

The Noisy Stochastic Gradient Descent Mechanism (NoisySGD)

- Utility analysis:
  - A composition of T subsampled gaussian mechanism.

Next lecture

- Differentially private deep learning
- Knowledge transfer model of private learning