Parameter-Free Locally Differentially Private Stochastic Subgradient Descent

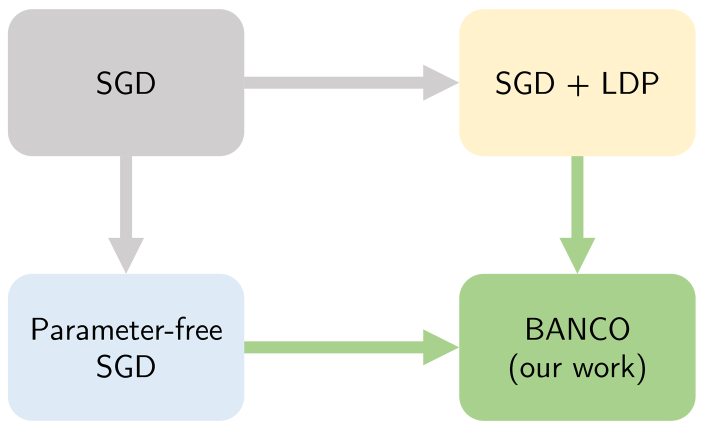
Kwang-Sung Jun (U of Arizona) and Francesco Orabona (Boston University)

Abstract

[The problem] Perform Stochastic subGradient Descent (SGD) while guaranteeing ϵ-locally differentially private (ϵ-LDP).

[The need] Convergence rates of existing solutions (e.g., (Song et al., 2013)) largely depend on the learning rate that must be *tuned* via repeated runs.
$\Longrightarrow$ privacy sacrificed!

[Goal] Converge as fast as the best learning rate in hindsight, in one pass & without tuning.



[Contribution] We propose BANCO (Betting Algorithm for Noisy COins), the first ϵ-LDP SGD algorithm that essentially matches the convergence rate of the tuned SGD without any learning rate parameter, reducing privacy loss and saving privacy budget.

[What do you mean by the best learning rate?]

$$w^* := \min_w R(w) \quad \text{where} \quad R(w) \text{ is the expected loss of } w \ .$$

The standard SGD with a constant learning rate η with ϵ-LDP guarantee:

$$\mathbb{E}[R(w_T)] - R(w^*) = O\left(\frac{\|w^*\|^2}{\eta T} + \frac{d^2}{\epsilon^2}\eta\right) \ .$$

The optimal rate would be

$$O\left(\frac{d}{\epsilon}\|w^*\|/\sqrt{T}\right) \quad \text{with} \quad \eta = \|w^*\|\frac{\epsilon/d}{\sqrt{T}}, \qquad \ldots??!! \text{ but who knows } \|w^*\|?$$

- In reality, the best bound is $O(\frac{d}{\epsilon}\|w^*\|^2/\sqrt{T})$ with $\eta = \frac{\epsilon/d}{\sqrt{T}}$.
- In practice, must tune the learning rate with *repeated runs*..

In contrast, BANCO achieves the rate $\frac{d}{\epsilon}\|w^*\|/\sqrt{T}$ up to logarithmic factors without knowing $\|w^*\|$!

[Why is parameter-free nontrivial for ϵ-LDP?] Existing techniques require the observed gradients to be bounded, but for LDP gradients are corrupted with *unbounded* noise.

Problem definition

We consider SGD for minimizing the **test loss** (rather than train loss) with access to **sanitized subgradients** (i.e., corrupted by noise).

- The loss $\ell(w, x)$: convex in w, and x is the sensitive data about an individual.
- The test loss $R(w) := \mathbb{E}_{x\sim\rho_X}[\ell(w,x)]$ where ρ_X is the distribution of the sensitive data.
- Sanitized subgradients: a noisy version $\mathcal{G}(w) \in \partial\ell(w,x) + \xi$ where the noise ξ guarantees the ϵ-LDP.
- Task: Perform SGD with sanitized subgradient requests; converge as close as possible to w^* after T iterations.

[Assumptions]

w_t: SGD iterate at time t, $\quad \hat{g}_t$: sanitized (negative) subgradient of w_t, $\quad \xi_t := \hat{g}_t - \mathbb{E}[\hat{g}_t]$.

- **(A1)** $\|\mathbb{E}[\hat{g}_t]\|_2 \le G, \forall t$.
- **(A2)** Bounded variance: $\mathbb{E}\left[\|\xi_t\|_2^2 \mid \xi_{1:t-1}\right] \le \sigma^2, \forall t$.
- **(A3)** Tail condition: $\xi_t|\xi_{1:t-1}$ is $(\sigma_{1\mathrm{D}}^2, b)$-sub-exponential

$$\max_{a:\|a\|_2\le 1} \mathbb{E}_t\left[\exp(\beta\langle\xi_t, a\rangle)\right] \le \exp\left(\frac{\beta^2\sigma_{1\mathrm{D}}^2}{2}\right), \ \forall|\beta| \le \frac{1}{b}$$

Definition 1 (Local Differential Privacy) *Let $D = (X_1, \ldots, X_n)$ be a sensitive dataset where each $X_i \sim \rho_X$ corresponds to data about individual i. A randomized sanitization mechanism M which outputs a disguised version $(U_1, \ldots U_n)$ of D is said to provide ϵ-local differential privacy to individual i, if*

$$\sup_S \sup_{x,x'\in\mathcal{D}} \frac{\mathbb{P}[U_i \in S|X_i = x]}{\mathbb{P}[U_i \in S|X_i = x']} \le \exp(\epsilon),$$

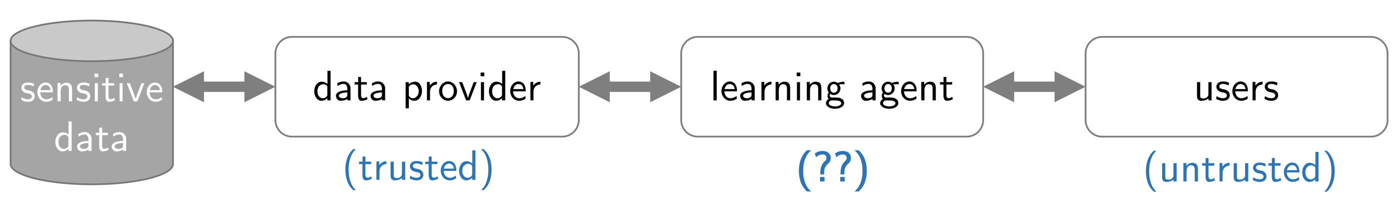
where the probability is w.r.t. the randomization in the sanitization mechanism.

[Example] The Laplace sanitization mechanism samples the noise ξ by

$$\rho_\xi(z) \propto \exp(-\frac{\epsilon}{2}\|z\|_2)$$

- Guarantees ϵ-LDP
- Satisfies **(A2)**: $\mathbb{E}\left[\|\xi_t\|_2^2\right] \le \frac{4(d^2+d)}{\epsilon^2}$; $\qquad$ **(A3)**: $\sigma_{1\mathrm{D}}^2 = 18d^2/\epsilon^2$ and $b = \epsilon/4$.

Related Work



- DP **vs** LDP: In DP, the data provider trusts the learning agent. LDP does not, so the data itself must be sanitized.
- Minimize empirical risk (ERM) **vs** true risk (generalization).
- Song et al. (2013; 2015): LDP, ERM.
- Wu et al. (2017): DP, ERM.
- Duchi et al. (2014); Bassily et al. (2014): LDP, generalization.

Those that tune the learning rate assume the bounded domain. $\Longrightarrow$ unrealistic and suboptimal.
$\Longrightarrow$ Ours: LDP, generalization, convergence rate of the tuned SGD without tuning!

Parameter-free stochastic optimization with noise

Algorithm 1 Betting Algorithm for Noisy COins (BANCO) for Locally Differentially Private SGD

Set $w_1 = q_1 = \mathbf{0} \in \mathbb{R}^d$
for $t = 1$ **to** T **do**
  Receive a noisy negative subgradient $\hat{g}_t$ such that $\mathbb{E}[\hat{g}_t] \in -\partial\ell(w_t, x)$ where $x \sim \rho_X$
  Update magnitude: $m_{t+1} = \frac{1}{2a}\int_{-a}^{a}\beta\exp\left(\beta\sum_{s=1}^t\langle\hat{g}_s, q_s\rangle - \beta^2 t\left(\frac{\sigma^2}{2} + G^2\right)\right)\mathrm{d}\beta$
  where $a = \min\left(\frac{k_1}{G}, \frac{1}{b}\right)$, and $k_1 = 0.6838$
  Update direction: $q_{t+\frac{1}{2}} = q_t - \frac{\hat{g}_t}{\sqrt{\sum_{s=1}^t\|\hat{g}_s\|_2^2}}$
  Project direction onto $L2$ ball: $q_{t+1} = q_{t+\frac{1}{2}} \cdot \min\left(1, \left\|q_{t+\frac{1}{2}}\right\|_2^{-1}\right)$
  Update the weight vector: $w_{t+1} = m_{t+1}q_{t+1} \in \mathbb{R}^d$
end for
Return $\frac{1}{T}\sum_{t=1}^T w_t$

A closed form solution of m_{t+1}: with shorthands $x = \sum_{s=1}^t\langle\hat{g}_s, q_s\rangle$ and $y = t(\sigma^2/2 + G^2)$,

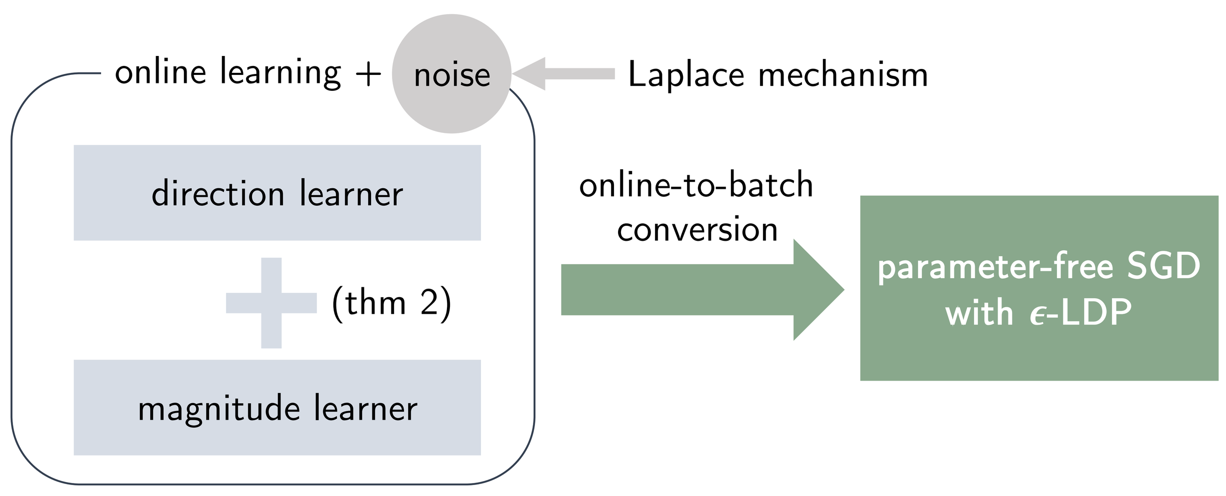
$$m_{t+1} = \frac{e^{-a(ay+x)}\left(\sqrt{\pi}x\exp\left(\frac{(2ay+x)^2}{4y}\right)\left(\mathrm{erf}(\frac{2ay+x}{2\sqrt{y}}) + \mathrm{erf}(\frac{2ay-x}{2\sqrt{y}})\right) + 2\sqrt{y}(1-e^{2ax})\right)}{8ay^{3/2}} \ .$$

Theorem 1 *Let $G = 1$. Let the noise ξ_t follow the Laplace mechanism. Then, for any $w^\star \in \mathbb{R}^d$, after one pass over T samples Algorithm 1 guarantees*

$$\mathbb{E}\left[R\left(\frac{1}{T}\sum_{t=1}^T w_t\right)\right] - R(w^\star) \le O\left(\frac{d\|w^\star\|_2}{\epsilon\sqrt{T}}\sqrt{\ln\left(1 + \frac{d^2\|w^\star\|_2 T}{\epsilon^2}\right)} + \frac{1}{T}\right) \ .$$

- Unimprovable up to logarithmic factors (Jun and Orabona, 2019).
- A more general version in (Jun and Orabona, 2019): extension to Banach space, connection to concentration inequalities, etc.
- The Gaussian noise can also be used, resulting a better dependency in the dimension of the space, but in the weaker (ϵ, δ)-LDP.

Proof Sketch



Key: The flexibility of "regret" in online learning allows combining two learners.
Assume:

- Direction: $R_T^{\mathrm{D}}(u) := \mathbb{E}[\sum_{t=1}^T\langle\hat{g}_t, u - q_t\rangle], \forall u : \|u\|_2 \le 1$.
- Magnitude: $R_T^{\mathrm{M}}(v) := \mathbb{E}[\sum_{t=1}^T s_t \cdot (v - m_t)]$ where $s_t = \langle\hat{g}_t, q_t\rangle, \forall v \in \mathbb{R}$.

Theorem 2 *Let $g_t := \mathbb{E}[\hat{g}_t]$. The iterates $m_t q_t$ guarantee, $\forall u \in \mathbb{R}^d$,*

$$\mathbb{E}\,\mathrm{Regret}_T(u) := \mathbb{E}\sum_{t=1}^T\langle g_t, u - m_t q_t\rangle \le R_T^{\mathrm{M}}(\|u\|) + \|u\|R_T^{\mathrm{D}}(\tfrac{u}{\|u\|}) \ .$$

- Direction learner: projected online gradient descent with the scale-free learning rates.

$$\mathbb{E}\left[R_T^{\mathrm{D}}\left(\frac{u}{\|u\|_2}\right)\right] = O\left(\mathbb{E}\left[\sqrt{\sum_{t=1}^T\|\hat{g}_t\|_2^2}\right]\right) \overset{(a)}{=} O\left(\sqrt{\sum_{t=1}^T(\mathbb{E}\|g_t\|_2^2 + \sigma^2)}\right),$$

  where (a) uses Jensen's inequality and the fact that $\mathbb{E}[\|\hat{g}_t\|_2^2] = \mathbb{E}[\|g_t\|_2^2] + \sigma^2$.
- Magnitude learner: the coin betting algorithm of Jun and Orabona (2019) that enjoys:

$$R_T^{\mathrm{D}}(u) = O\left(|u|\max\left\{(1+b)\ln(|u|(1+b)), \sqrt{(1+\sigma_{1\mathrm{D}}^2)T\ln(|u|(1+\sigma_{1\mathrm{D}}^2)T+1)}\right\} + 1\right) \ .$$

Future work

- High probability convergence guarantees.
- Through empirical evaluation of BANCO.
- ~~Data-dependent regret bound that depends on $\|\hat{g}_t\|^2$ rather than $(G^2 + \sigma^2)T$.~~
- ~~Be agnostic to the noise parameters (σ^2, b).~~

The last two are resolved by a followup paper by van der Hoeven (2019) for symmetric noise.

References

R. Bassily, A. Smith, and A. Thakurta. Differentially private empirical risk minimization: Efficient algorithms and tight error bounds. *arXiv preprint arXiv:1405.7085*, 2014.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Privacy aware learning. *Journal of the ACM*, 61(6):38, 2014.

K.-S. Jun and F. Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Proc. of the Conference on Learning Theory (COLT)*, 2019.

S. Song, K. Chaudhuri, and A. D. Sarwate. Stochastic gradient descent with differentially private updates. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 245–248. IEEE, 2013.

S. Song, K. Chaudhuri, and A. Sarwate. Learning from data with heterogeneous noise using SGD. In *Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 894–902, 2015.

D. van der Hoeven. User-specified local differential privacy in unconstrained adaptive online learning. In *Advances in Neural Information Processing Systems 32*, pages 14080–14089. 2019.

X. Wu, F. Li, A. Kumar, K. Chaudhuri, S. Jha, and J. Naughton. Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proc. of the 2017 ACM International Conference on Management of Data*, pages 1307–1322. ACM, 2017.