Differentially Private Stochastic Gradient Descent with Fixed-Size Minibatches: Tighter RDP Guarantees with or without Replacement

Jeremiah Birrell (Texas State University)
Reza Ebrahimi (University of South Florida)
Rouzbeh Behnia (University of South Florida)
Jason Pacheco (University of Arizona)

NeurIPS 2024

Data Privacy violations in Machine Learning

Privacy leakage: Machine learning (ML) models are known to be susceptible to privacy leakage attacks. Information about the training set can often be extracted from a trained model by an attacker. Notably, sensitive data can be obtained by membership inference attacks[1].

This is a big problem if the data is sensitive and the model is public.

Differential Privacy:

- Facilitates provably private training of ML models.
- Modern formulation built on the pioneering work of Cynthia Dwork et al.[2]

[1] R. Shokri, M. Stronati, C. Song, V. Shmatikov, 2017 IEEE Symposium on Security and Privacy.

[2] C. Dwork, F. McSherry, K. Nissim and A. D. Smith, Theory of Cryptography, (2006).

Data Privacy violations in Machine Learning

Privacy leakage: Machine learning (ML) models are known to be susceptible to privacy leakage attacks. Information about the training set can often be extracted from a trained model by an attacker. Notably, sensitive data can be obtained by membership inference attacks[1].

This is a big problem if the data is sensitive and the model is public.

Differential Privacy:

- Facilitates provably private training of ML models.
- Modern formulation built on the pioneering work of Cynthia Dwork et al.[2]

[1] R. Shokri, M. Stronati, C. Song, V. Shmatikov, 2017 IEEE Symposium on Security and Privacy.

[2] C. Dwork, F. McSherry, K. Nissim and A. D. Smith, Theory of Cryptography, (2006).

Data Privacy violations in Machine Learning

Privacy leakage: Machine learning (ML) models are known to be susceptible to privacy leakage attacks. Information about the training set can often be extracted from a trained model by an attacker. Notably, sensitive data can be obtained by membership inference attacks[1].

This is a big problem if the data is sensitive and the model is public.

Differential Privacy:

- Facilitates provably private training of ML models.
- Modern formulation built on the pioneering work of Cynthia Dwork et al.[2]

[1] R. Shokri, M. Stronati, C. Song, V. Shmatikov, 2017 IEEE Symposium on Security and Privacy.

[2] C. Dwork, F. McSherry, K. Nissim and A. D. Smith, Theory of Cryptography, (2006).

Data Privacy violations in Machine Learning

Privacy leakage: Machine learning (ML) models are known to be susceptible to privacy leakage attacks. Information about the training set can often be extracted from a trained model by an attacker. Notably, sensitive data can be obtained by membership inference attacks[1].

This is a big problem if the data is sensitive and the model is public.

Differential Privacy:

- Facilitates provably private training of ML models.
- Modern formulation built on the pioneering work of Cynthia Dwork et al.[2]

[1] R. Shokri, M. Stronati, C. Song, V. Shmatikov, 2017 IEEE Symposium on Security and Privacy.

[2] C. Dwork, F. McSherry, K. Nissim and A. D. Smith, Theory of Cryptography, (2006).

Data Privacy violations in Machine Learning

Privacy leakage: Machine learning (ML) models are known to be susceptible to privacy leakage attacks. Information about the training set can often be extracted from a trained model by an attacker. Notably, sensitive data can be obtained by membership inference attacks[1].

This is a big problem if the data is sensitive and the model is public.

Differential Privacy:

- Facilitates provably private training of ML models.
- Modern formulation built on the pioneering work of Cynthia Dwork et al.[2]

[1] R. Shokri, M. Stronati, C. Song, V. Shmatikov, 2017 IEEE Symposium on Security and Privacy.

[2] C. Dwork, F. McSherry, K. Nissim and A. D. Smith, Theory of Cryptography, (2006).

Data Privacy violations in Machine Learning

Privacy leakage: Machine learning (ML) models are known to be susceptible to privacy leakage attacks. Information about the training set can often be extracted from a trained model by an attacker. Notably, sensitive data can be obtained by membership inference attacks[1].

This is a big problem if the data is sensitive and the model is public.

Differential Privacy:

- Facilitates provably private training of ML models.
- Modern formulation built on the pioneering work of Cynthia Dwork et al.[2]

---

[1] R. Shokri, M. Stronati, C. Song, V. Shmatikov, 2017 IEEE Symposium on Security and Privacy.

[2] C. Dwork, F. McSherry, K. Nissim and A. D. Smith, Theory of Cryptography, (2006).

Rényi-DP

Rényi-DP[1]**:** A random algorithm $\mathcal{M}$ has (ϵ, α)-RDP under the dataset adjacency relation $\simeq$ if

$$\sup_{D,D':D\simeq D'} R_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) \leq \epsilon$$

Rényi Divergences of order $\alpha > 1$:

$$R_\alpha(Q\|P) := \frac{1}{\alpha - 1} \log E_P[(dQ/dP)^\alpha]$$

[1] I. Mironov, Proc. IEEE Comp. Security Foundations Symp. (CSF), (2017)

Rényi-DP

Rényi-DP[1]: A random algorithm $\mathcal{M}$ has (ϵ, α)-RDP under the dataset adjacency relation $\simeq$ if

$$\sup_{D,D':D\simeq D'} R_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) \leq \epsilon$$

Rényi Divergences of order $\alpha > 1$:

$$R_\alpha(Q\|P) := \frac{1}{\alpha - 1} \log E_P[(dQ/dP)^\alpha]$$

[1] I. Mironov, Proc. IEEE Comp. Security Foundations Symp. (CSF), (2017)

RDP Bounds for SGD with Poisson Subsampling

Differentially private SGD:

$$\theta_{t+1} = \theta_t - \eta_t \frac{1}{|B|} \sum_{d \in B_t} \mathrm{Clip}[\nabla \mathcal{L}_d(\theta_t)] + \sigma_n N(0, I)$$

There are multiple ways to form minibatches, B_t.

Poisson subsampling: Minibatches are formed by iid Bernoulli random variables (chosen sampling probability q) which decide whether each sample is included in the minibatch or not.

RDP bounds on SGD with Poisson subsampling: First bounds obtained by Abadi et al.[1] and Mironov et al.[2]

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L., Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, (2016)

[2] Mironov, I., Talwar, K., and Zhang, L., arXiv:1908.10530, (2019).

RDP Bounds for SGD with Poisson Subsampling

Differentially private SGD:

$$\theta_{t+1} = \theta_t - \eta_t \frac{1}{|B|} \sum_{d \in B_t} \mathrm{Clip}[\nabla \mathcal{L}_d(\theta_t)] + \sigma_n N(0, I)$$

There are multiple ways to form minibatches, B_t.

Poisson subsampling: Minibatches are formed by iid Bernoulli random variables (chosen sampling probability q) which decide whether each sample is included in the minibatch or not.

RDP bounds on SGD with Poisson subsampling: First bounds obtained by Abadi et al.[1] and Mironov et al.[2]

---

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L., Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, (2016)

[2] Mironov, I., Talwar, K., and Zhang, L., arXiv:1908.10530, (2019).

RDP Bounds for SGD with Poisson Subsampling

Differentially private SGD:

$$\theta_{t+1} = \theta_t - \eta_t \frac{1}{|B|} \sum_{d \in B_t} \text{Clip}[\nabla \mathcal{L}_d(\theta_t)] + \sigma_n N(0, I)$$

There are multiple ways to form minibatches, B_t.

Poisson subsampling: Minibatches are formed by iid Bernoulli random variables (chosen sampling probability q) which decide whether each sample is included in the minibatch or not.

RDP bounds on SGD with Poisson subsampling: First bounds obtained by Abadi et al.[1] and Mironov et al.[2]

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L., Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, (2016)

[2] Mironov, I., Talwar, K., and Zhang, L., arXiv:1908.10530, (2019).

RDP Bounds for SGD with Poisson Subsampling

Differentially private SGD:

$$\theta_{t+1} = \theta_t - \eta_t \frac{1}{|B|} \sum_{d \in B_t} \mathrm{Clip}[\nabla \mathcal{L}_d(\theta_t)] + \sigma_n N(0, I)$$

There are multiple ways to form minibatches, B_t.

Poisson subsampling: Minibatches are formed by iid Bernoulli random variables (chosen sampling probability q) which decide whether each sample is included in the minibatch or not.

RDP bounds on SGD with Poisson subsampling: First bounds obtained by Abadi et al.[1] and Mironov et al.[2]

[1] **Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L., Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, (2016)**

[2] **Mironov, I., Talwar, K., and Zhang, L., arXiv:1908.10530, (2019).**

RDP for Fixed-size Subsampling without Replacement

Disadvantage of Poisson subsampling: Leads to variable sized minibatches and therefore inconsistent memory usage. It also has higher variance.

Fixed-size subsampling: Constant memory usage, but RDP bounds more difficult to obtain.

General purpose RDP bounds (i.e., for general $\mathcal{M}$) with fixed-size subsampling obtained by Wang et al.[1]

We obtain **tighter RDP bounds** for fixed-size subsampled DP-SGD using a Taylor expansion method, with precise bounds on the expansion remainder terms [2].

---

[1] Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)

[2] Birrell, J., Ebrahimi, R., Behnia, R., Pacheco, J., NeurIPS (2024), arXiv:2408.10456.

RDP for Fixed-size Subsampling without Replacement

Disadvantage of Poisson subsampling: Leads to variable sized minibatches and therefore inconsistent memory usage. It also has higher variance.

Fixed-size subsampling: Constant memory usage, but RDP bounds more difficult to obtain.

General purpose RDP bounds (i.e., for general $\mathcal{M}$) with fixed-size subsampling obtained by Wang et al.[1]

We obtain **tighter RDP bounds** for fixed-size subsampled DP-SGD using a Taylor expansion method, with precise bounds on the expansion remainder terms [2].

[1] Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)

[2] Birrell, J., Ebrahimi, R., Behnia, R., Pacheco, J., NeurIPS (2024), arXiv:2408.10456.

RDP for Fixed-size Subsampling without Replacement

Disadvantage of Poisson subsampling: Leads to variable sized minibatches and therefore inconsistent memory usage. It also has higher variance.

Fixed-size subsampling: Constant memory usage, but RDP bounds more difficult to obtain.

General purpose RDP bounds (i.e., for general $\mathcal{M}$) with fixed-size subsampling obtained by Wang et al.[1]

We obtain **tighter RDP bounds** for fixed-size subsampled DP-SGD using a Taylor expansion method, with precise bounds on the expansion remainder terms [2].

[1] Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)

[2] Birrell, J., Ebrahimi, R., Behnia, R., Pacheco, J., NeurIPS (2024), arXiv:2408.10456.

RDP for Fixed-size Subsampling without Replacement

Disadvantage of Poisson subsampling: Leads to variable sized minibatches and therefore inconsistent memory usage. It also has higher variance.

Fixed-size subsampling: Constant memory usage, but RDP bounds more difficult to obtain.

General purpose RDP bounds (i.e., for general $\mathcal{M}$) with fixed-size subsampling obtained by Wang et al.[1]

We obtain **tighter RDP bounds** for fixed-size subsampled DP-SGD using a Taylor expansion method, with precise bounds on the expansion remainder terms [2].

[1] **Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)**

[2] **Birrell, J., Ebrahimi, R., Behnia, R., Pacheco, J., NeurIPS (2024), arXiv:2408.10456.**

RDP for Fixed-size Subsampling without Replacement

Disadvantage of Poisson subsampling: Leads to variable sized minibatches and therefore inconsistent memory usage. It also has higher variance.

Fixed-size subsampling: Constant memory usage, but RDP bounds more difficult to obtain.

General purpose RDP bounds (i.e., for general $\mathcal{M}$) with fixed-size subsampling obtained by Wang et al.[1]

We obtain **tighter RDP bounds** for fixed-size subsampled DP-SGD using a Taylor expansion method, with precise bounds on the expansion remainder terms [2].

---

[1] Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)

[2] Birrell, J., Ebrahimi, R., Behnia, R., Pacheco, J., NeurIPS (2024), arXiv:2408.10456.

RDP SGD under Fixed-size Subsampling

Theorem (T-step FS_{woR}-RDP Upper Bound under Replace-one Adjacency[1])

Assuming $q < 1$ (sampling probability), T-step fixed-size subsampled (without replacement) DP-SGD has $(\alpha, \epsilon_{[0,T]}(\alpha))$-RDP under replace-one adjacency, where

$$\epsilon_{[0,T]}(\alpha) \le \sum_{t=0}^{T-1} \frac{1}{\alpha - 1} \log \left[1 + q^2 \alpha(\alpha - 1) \left(e^{4/\sigma_t^2} - e^{2/\sigma_t^2} \right) + O(q^3) \right]$$

1. We provide computatable bounds on the $O(q^3)$ term.
2. Our result improves on the RDP bound of Wang et al.[2] by approximately a factor of 4 and is close to the theoretical lower bound[2] in practice.

[1] Birrell, J., Ebrahimi, R., Behnia, R., Pacheco, J., NeurIPS (2024), arXiv:2408.10456.

[2] Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)

RDP SGD under Fixed-size Subsampling

Theorem (T-step FS_{woR}-RDP Upper Bound under Replace-one Adjacency[1])

Assuming $q < 1$ (sampling probability), T-step fixed-size subsampled (without replacement) DP-SGD has $(\alpha, \epsilon_{[0,T]}(\alpha))$-RDP under replace-one adjacency, where

$$\epsilon_{[0,T]}(\alpha) \leq \sum_{t=0}^{T-1} \frac{1}{\alpha - 1} \log \left[1 + q^2 \alpha(\alpha - 1) \left(e^{4/{\sigma_t}^2} - e^{2/{\sigma_t}^2} \right) + O(q^3) \right]$$

1. We provide computatable bounds on the $O(q^3)$ term.
2. Our result improves on the RDP bound of Wang et al.[2] by approximately a factor of 4 and is close to the theoretical lower bound[2] in practice.

[1] Birrell, J., Ebrahimi, R., Behnia, R., Pacheco, J., NeurIPS (2024), arXiv:2408.10456.

[2] Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)

RDP SGD under Fixed-size Subsampling

Theorem (T-step FS_{woR}-RDP Upper Bound under Replace-one Adjacency[1])

Assuming $q < 1$ (sampling probability), T-step fixed-size subsampled (without replacement) DP-SGD has $(\alpha, \epsilon_{[0,T]}(\alpha))$-RDP under replace-one adjacency, where

$$\epsilon_{[0,T]}(\alpha) \leq \sum_{t=0}^{T-1} \frac{1}{\alpha - 1} \log \left[1 + q^2 \alpha(\alpha - 1) \left(e^{4/\sigma_t^2} - e^{2/\sigma_t^2} \right) + O(q^3) \right]$$

1. We provide computatable bounds on the $O(q^3)$ term.
2. Our result improves on the RDP bound of Wang et al.[2] by approximately a factor of 4 and is close to the theoretical lower bound[2] in practice.

[1] Birrell, J., Ebrahimi, R., Behnia, R., Pacheco, J., NeurIPS (2024), arXiv:2408.10456.

[2] Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)

RDP SGD under Fixed-size Subsampling

Theorem (T-step FS_{woR}-RDP Upper Bound under Replace-one Adjacency[1])

Assuming $q < 1$ (sampling probability), T-step fixed-size subsampled (without replacement) DP-SGD has $(\alpha, \epsilon_{[0,T]}(\alpha))$-RDP under replace-one adjacency, where

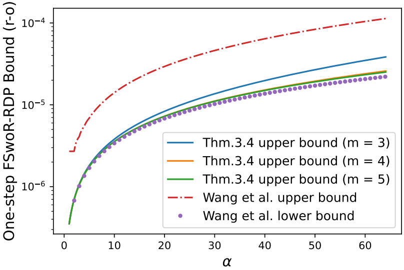
$$\epsilon_{[0,T]}(\alpha) \leq \sum_{t=0}^{T-1} \frac{1}{\alpha - 1} \log \left[1 + q^2 \alpha(\alpha - 1) \left(e^{4/\sigma_t{}^2} - e^{2/\sigma_t{}^2} \right) + O(q^3) \right]$$

1. We provide computatable bounds on the $O(q^3)$ term.
2. Our result improves on the RDP bound of Wang et al.[2] by approximately a factor of 4 and is close to the theoretical lower bound[2] in practice.

[1] Birrell, J., Ebrahimi, R., Behnia, R., Pacheco, J., NeurIPS (2024), arXiv:2408.10456.

[2] Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)

Comparison with Wang et al.



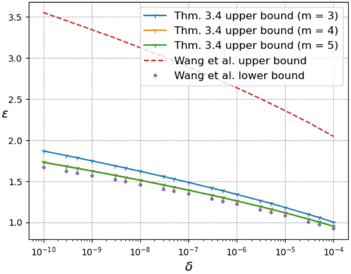


Figure: Comparison of our FS_{woR}-RDP bounds under replace-one adjacency for various choices of m with the upper and lower bounds from Wang et al.[1] We used $\sigma_t = 6$, $|B| = 120$, and $|D| = 50,000$.

[1] Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P., PMLR, (2019)

Comparison with Poisson Subsampling on CIFAR10

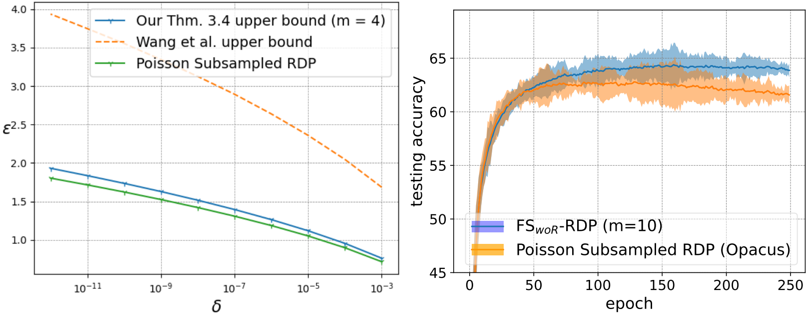


Figure: Comparing privacy guarantees of FS_{woR}-RDP with Wang et al. and Poisson Subsampled RDP (**left**). Comparing FS_{woR}-RDP performance against Poisson subsampled RDP (**right**). We used $\sigma_t = 6$, $C = 3$, $|B| = 120$, $|D| = 50,000$, and $lr = 1e-3$.

Memory Usage Comparison

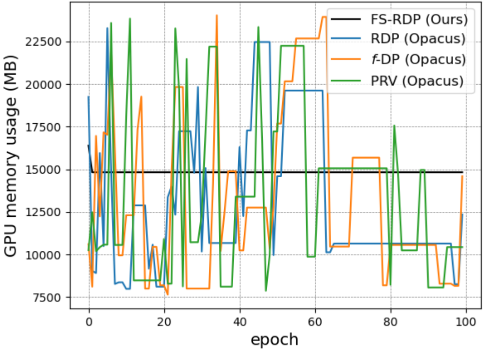


Figure: Comparing memory usage of FS-RDP with other Opacus privacy accountants in each training epoch. We used $|B| = 120$, and $|D| = 50,000$.

For further details see:
Differentially Private Stochastic Gradient Descent with Fixed-Size Minibatches: Tighter RDP Guarantees with or without Replacement, J. Birrell, R. Ebrahimi, R. Behnia, J. Pacheco, NeurIPS (2024)

Preprint: arXiv:2408.10456