Equivariant Differentially Private Deep Learning: Exploiting Symmetry under Privacy

Margarita Ionides [1]

Abstract

Differentially private stochastic gradient descent (DP-SGD) is the current standard for training neural networks under formal privacy guarantees, but its accuracy degrades sharply in high-dimensional parameter spaces. Equivariant neural networks can mitigate some of the accuracy loss caused by DP noise, in part by reducing parameter redundancy through symmetry. However, DP-SGD fails to take symmetry into account: in standard DP-SGD, isotropic Gaussian noise is injected uniformly in parameter space and gradients are clipped in the ℓ_2 norm, ignoring the geometry of the equivariant parameterization. We introduce *representation-aligned* DP-SGD, which aligns both gradient clipping and noise injection with the equivariant parameterization while preserving the same (ε, δ) privacy guarantees of the standard method. Empirically, we find that representation-aligned noise significantly improves upon isotropic baselines under matched privacy budgets.

1. Introduction

Differential privacy (DP) (Dwork & Roth, 2014; Dwork, 2006; Dwork et al., 2006) provides strong guarantees that trained models do not reveal information about individual training examples. In deep learning, the dominant approach is differentially private stochastic gradient descent (DP-SGD) (Abadi et al., 2016), which enforces privacy by clipping per-sample gradients and injecting Gaussian noise. While effective in privacy, DP-SGD often incurs substantial accuracy degradation, particularly in high-dimensional vision models (Hölzl et al., 2023; Tramèr & Boneh, 2021; Duan et al., 2025).

Group-equivariant neural networks, a central tool in geometric deep learning, incorporate known symmetries directly into model architectures (Cohen & Welling, 2016; Weiler & Cesa, 2019; Cesa et al., 2022). These models reduce parameter redundancy and often improve efficiency and robustness in equivariant tasks, and in particular vision tasks (Bronstein et al., 2021). Recent work suggests that equivariant architectures can also mitigate some of the utility loss induced by DP-SGD (Hölzl et al., 2023).

However, the DP mechanism itself remains oblivious to this structure. Standard DP-SGD clips gradients in the Euclidean norm and injects isotropic noise, treating all parameter directions as equally sensitive. In short, we know (Hölzl et al., 2023) that choosing a model that respects the geometric structure of its data leads to improved outcomes in DP training, and yet DP mechanisms are not taking advantage of this structure. This raises a natural question: can the DP-SGD mechanism itself be redesigned to respect the representation structure of equivariant networks?

We propose **representation-aligned DP-SGD**, a symmetry-aware DP-SGD. The key innovation of representation-aligned DP-SGD is to align both gradient clipping and noise injection with a network's equivariant parameterization. Our approach leverages the steerable kernel basis used in equivariant CNNs (Weiler & Cesa, 2019; Cesa et al., 2022) together with its associated generalized He-initialization (He et al., 2015), which provides closed-form variance scales for each basis element. By whitening gradients using these scales and performing clipping and Gaussian noise addition in the whitened coordinate system, we obtain a DP mechanism whose noise geometry matches the model geometry while preserving the privacy guarantees of DP-SGD. Unlike recent geometric perspectives on DP-SGD, which have sought to decouple directional and magnitude noise to improve convergence (Duan et al., 2025), or prior improvements to DP-SGD that focus on clipping heuristics, optimization schedules, or adaptive noise scaling (Thakkar et al., 2019; Yu et al., 2021; Pichapati et al., 2019), our approach instead targets the representation-theoretic geometry of the model itself.

Empirically, we find that representation-aligned DP-SGD improves test accuracy over isotropic baselines across a range of noise levels on rotation-equivariant vision tasks

[1]Department of Statistics, University of Michigan, Ann Arbor, MI, USA. Correspondence to: Margarita Ionides <rionides@umich.edu>.

under matched privacy budgets. We further test diagnostic ablations that deliberately misalign the noise geometry of the DP mechanism, demonstrating that the observed gains arise from respecting the equivariant parameterization rather than from architectural or optimization artifacts. Together, these results suggest that respecting representation geometry is an important consideration when designing differentially private mechanisms for equivariant models.

2. Representation-Aligned DP-SGD

2.1. Differential Privacy

We follow standard differentially private stochastic gradient descent (Abadi et al., 2016), which achieves (ε, δ)-differential privacy (Dwork & Roth, 2014) by clipping per-sample gradients and adding Gaussian noise calibrated to the clipping bound. Privacy loss is tracked using Rényi differential privacy (Mironov, 2017).

2.2. Equivariant Parameterization and Geometry

Group-equivariant convolutional networks (Weiler & Cesa, 2019; Cesa et al., 2022) constrain convolutional kernels to lie in a linear subspace determined by group representations. In `escnn`, kernels are parameterized using an orthonormal steerable basis $\{B_i\}_{i=1}^k$:

$$K(\theta) = \sum_{i=1}^{k} \theta_i B_i, \quad \theta \in \mathbb{R}^k,$$

where the coefficients $\theta = (\theta_1, \dots, \theta_k)$ are the learnable parameters. Because each basis element corresponds to specific irreducible representations of the symmetry group, the coefficients θ_i can differ substantially in scale and functional impact.

To account for this heterogeneity, `escnn` provides generalized He-initialization scales (He et al., 2015), defined by

$$m_i = \sqrt{\frac{2}{n_{in}\, \mathbb{E}_x[\|B_i * x\|_2^2]}}$$

where n_{in} denotes the input dimensionality, and the expectation is calculated over random unit-variance inputs. The diagonal matrix $M = \mathrm{diag}(m_1, \dots, m_k)$ defines a representation-induced metric on parameter space, which captures the difference in sensitivity across basis directions.

This motivates our representation-aligned approach: rather than applying the isotropic ℓ_2-clipping standard in DP-SGD, we clip gradients in the whitened coordinate system defined by M, ensuring that privacy-preserving noise is added proportional to the functional impact of each parameter. This metric is fixed prior to training, independent of the training data, and reflects the intrinsic geometry induced by the equivariant parameterization: we can therefore use it to define a symmetry-aware variant of DP-SGD.

2.3. Proposed Mechanism

Let $g \in \mathbb{R}^k$ denote the per-sample gradient with respect to the basis coefficients θ. Given a clipping threshold $C > 0$ and noise multiplier $\sigma > 0$, representation-aligned DP-SGD algorithm is presented in Algorithm 1. The mechanism is applied independently to each sample in a minibatch, and the resulting updates are averaged as in standard DP-SGD.

Crucially, clipping and noise injection are performed in the whitened coordinate system induced by M, while the final update is obtained via a deterministic linear transformation. As a result, the mechanism can be analyzed as a standard Gaussian mechanism in whitened space, followed by post-processing.

Algorithm 1 Representation-Aligned DP-SGD

Input: Dataset $\mathcal{D}$, model f_θ, symmetry group Γ, steerable basis $\{B_i\}_{i=1}^k$, epochs E, batch size B, learning rate η, noise multiplier σ, clipping threshold C.
Initialize: Compute $M = \mathrm{diag}(m_1, \dots, m_k)$ analytically from Γ and $\{B_i\}$. Initialize $\theta_0 \sim \mathcal{N}(0, M^2)$.
for $e = 1, \dots, E$ **do**
  Sample a random batch $B_t \subset \mathcal{D}$ of size B.
  for each sample $x_i \in B_t$ **do**
    Compute per-sample gradient,
    $g_i \leftarrow \nabla_\theta \mathcal{L}(f_\theta(x_i), y_i)$.
    $\tilde{g}_i \leftarrow M^{-1} g_i$ {Whiten gradient}
    $\tilde{g}_{i,\text{clip}} \leftarrow \tilde{g}_i \cdot \min(1, \frac{C}{\|\tilde{g}_i\|_2})$ {Gradient clipping}
  end for
  Sample noise $\tilde{Z} \sim \mathcal{N}(0, \sigma^2 C^2 I_k)$.
  $\tilde{G}_t \leftarrow \frac{1}{B}\left(\sum_{i=1}^{B} \tilde{g}_{i,\text{clip}} + \tilde{Z}\right)$.
  $\Delta\theta_t \leftarrow M\tilde{G}_t$ {Unwhiten}
  $\theta_t \leftarrow \theta_{t-1} - \eta \Delta\theta_t$
end for
Return: Private parameters θ_T.

2.4. Privacy Guarantee

We now show that representation-aligned DP-SGD satisfies the same privacy guarantees as the standard isotropic DP-SGD.

Proposition 2.1 (Privacy guarantee)**.** *The representation-aligned DP-SGD mechanism satisfies Rényi differential privacy with parameter* $\rho(\alpha) = \frac{\alpha}{2\sigma^2}$*. Consequently, after subsampling and composition, the mechanism satisfies a* (ε, δ)*-differential privacy guarantee.*

Proof sketch. Consider the intermediate mechanism operating in the whitened coordinate system:

$$\tilde{g}_{\text{priv}} = \tilde{g}_{\text{clip}} + \tilde{Z}, \qquad \tilde{Z} \sim \mathcal{N}(0, \sigma^2 C^2 I_k). \tag{1}$$

By construction, $\|\tilde{g}_{\text{clip}}\|_2 \leq C$, so the ℓ_2-sensitivity of the clipped gradient is exactly C. Therefore, the mechanism in whitened space is a standard Gaussian mechanism and satisfies Rényi differential privacy with

$$\rho(\alpha) = \frac{\alpha C^2}{2\sigma^2 C^2} = \frac{\alpha}{2\sigma^2}.$$

The final update $\Delta\theta = M\tilde{g}_{\text{priv}}$ is obtained by applying a deterministic, data-independent linear transformation. By the post-processing property of differential privacy, this transformation does not affect the privacy guarantee. Standard subsampling amplification and composition arguments then yield the same (ε, δ) guarantee as isotropic DP-SGD; full details are provided in Appendix A.

3. Experimental Setup

We train steerable E(2)-equivariant CNNs implemented in `escnn` (Cesa et al., 2022), comparing representation-aligned DP-SGD against isotropic DP-SGD under matched privacy budgets. All models share identical architectures and optimization settings; differences in performance therefore isolate the effect of noise geometry.

As an ablation, we also evaluate a shuffled-metric variant in which the diagonal scaling matrix M is randomly permuted across basis coefficients. This preserves the marginal noise distribution while deliberately misaligning noise with the equivariant parameterization.

We evaluate on RotMNIST and Rotated Fashion-MNIST, where images are randomly rotated by multiples of $45°$. Privacy loss is tracked using Rényi differential privacy accounting, and performance is evaluated using test-set classification accuracy.

4. Results

Figures 1 and 2 compare test accuracy as a function of privacy loss ε for isotropic (ISO) and representation-aligned (REP) DP-SGD on RotMNIST and Rotated Fashion-MNIST, respectively. Representation-aligned DP-SGD consistently improves accuracy over isotropic DP-SGD at matched privacy levels, with the largest gains in the moderate-privacy regime.

4.1. Ablation: Shuffled Representation Metric

To isolate the role of alignment from anisotropy alone, we evaluate a shuffled-metric ablation in which the diagonal scaling matrix M is randomly permuted across basis coefficients.

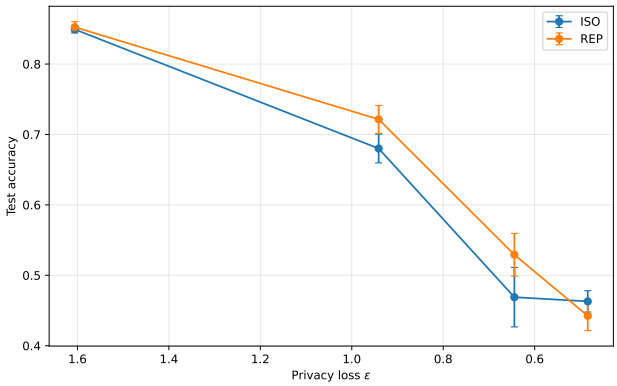

Figure 1. Utility-privacy trade-off on RotMNIST for isotropic (ISO) and representation-aligned (REP) DP-SGD. Representation-aligned DP-SGD consistently achieves higher accuracy at matched privacy levels, with the largest gains in the moderate-privacy regime.

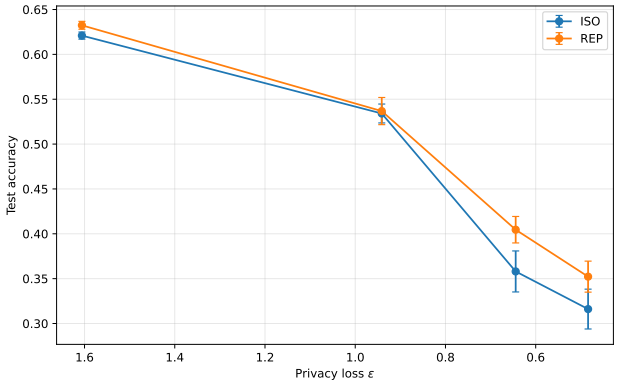

Figure 2. Utility–privacy trade-off on Rotated Fashion-MNIST. Representation-aligned DP-SGD achieves gains here too, with the largest gains occuring under high privacy.

Despite identical marginal noise statistics, the shuffled variant is less stable and typically underperforms the aligned mechanism. This indicates that the gains in representation-aligned DP-SGD arise from respecting the equivariant parameterization rather than from anisotropy alone.

5. Conclusion

We introduced representation-aligned DP-SGD, a symmetry-aware variant of DP-SGD that aligns gradient clipping and noise injection with the equivariant parameterization of group-equivariant neural networks. By leveraging generalized He-initialization scales, our method preserves the (ε, δ) privacy guarantee of DP-SGD while aligning its noise geometry with an equivariant model's structure. Empirically, we demonstrate consistent utility improvements under matched

Table 1. RotMNIST test accuracy for isotropic (ISO), shuffled-metric (Shuffle), and representation-aligned (REP) DP-SGD.Accuracy is reported as mean ± standard error across multiple random seeds.

σ	ε	ISO	Shuffle	REP
1.4	0.48	**46.3 ± 1.3**	45.7 ± 1.5	44.2 ± 1.8
1.2	0.64	46.9 ± 3.5	51.8 ± 3.7	**52.9 ± 2.5**
1.0	0.94	68.0 ± 1.8	60.9 ± 3.7	**72.1 ± 1.7**
0.8	1.61	84.9 ± 0.4	84.9 ± 0.5	**85.2 ± 0.7**
0.6	3.97	**89.9 ± 0.2**	89.7 ± 0.2	**89.9 ± 0.2**

privacy budgets, and show through our metric ablation study that alignment is truly responsible for these efficiency gains. These results identify noise geometry as an important and underexplored design dimension in differentially private learning.

Acknowledgements

The author thanks Amrita Roy Chowdhury and Ambuj Tewari for their mentorship and for fostering an environment encouraging independent exploration in differential privacy. The author is also grateful to the OpenDP group and Maani Ghaffari for discussions that influenced the development of this work. Special thanks to Saptarshi Roy for valuable discussions and careful proofreading.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

Amari, S.-I. *Information Geometry and Its Applications*. Springer, 2016.

Bronstein, M. M., Bruna, J., Cohen, T., and Velickovic, P. Geometric Deep Learning: Grids, groups, graphs, geodesics, and gauges. *arXiv:2104.13478*, 2021.

Cesa, G., Lang, L., and Weiler, M. A program to build E(N)-equivariant steerable CNNs. In *International Conference on Learning Representations*, 2022.

Cohen, T. and Welling, M. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pp. 2990–2999, 2016.

Duan, J., Hu, H., Ye, Q., and Sun, X. Analyzing and Optimizing Perturbation of DP-SGD Geometrically. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 3439–3452, Los Alamitos, CA, USA, May 2025. IEEE Computer Society. doi: 10.1109/ICDE65448.2025.00257. URL https://doi.ieeecomputersociety.org/10.1109/ICDE65448.2025.00257.

Dwork, C. Differential Privacy. In *Automata, languages and programming (ICALP 2006). Part II*, volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer, 2006.

Dwork, C. and Roth, A. *The Algorithmic Foundations of Differential Privacy*. Now Publishers, 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pp. 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *arXiv:1502.01852*, 2015.

Hölzl, F. A., Rueckert, D., and Kaissis, G. Equivariant differentially private deep learning: Why DP-SGD needs sparser models. *arXiv:2301.13104*, 2023.

Mironov, I. Rényi differential privacy. In *IEEE Computer Security Foundations Symposium*, pp. 263–275, 2017.

Pichapati, V., Suresh, A. T., Yu, F. X., Reddi, S. J., and Kumar, S. Adaclip: Adaptive clipping for private SGD. *CoRR*, abs/1908.07643, 2019. URL http://arxiv.org/abs/1908.07643.

Thakkar, O., Andrew, G., and McMahan, H. B. Differentially private learning with adaptive clipping. *CoRR*, abs/1905.03871, 2019. URL http://arxiv.org/abs/1905.03871.

Tramèr, F. and Boneh, D. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2021.

Weiler, M. and Cesa, G. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems*, pp. 14334–14345, 2019.

Weiler, M., Forré, P., Verlinde, E., and Welling, M. *Equivariant and Coordinate Independent Convolutional Networks: A Gauge Field Theory of Neural Networks*. World Scientific, 2023. doi: 10.1142/14143.

Yu, D., Zhang, H., Chen, W., and Liu, T. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. *CoRR*, abs/2102.12677, 2021. URL https://arxiv.org/abs/2102.12677.

A. Proof of Representation-Aligned DP-SGD

The proof relies on two main points: the sensitivity of the clipped whitened gradient and the Post-Processing Property of differential privacy.

We provide a complete proof that representation-aligned DP-SGD satisfies the same Rényi differential privacy (RDP) guarantees as standard isotropic DP-SGD. The argument proceeds in three steps: (i) we show that clipping in whitened coordinates induces a bounded ℓ_2-sensitivity query; (ii) we apply the Gaussian mechanism in whitened space and compute its RDP parameter; and (iii) we invoke the post-processing property of differential privacy to show that unwhitening does not affect the privacy guarantee.

Proof. Fix a training iteration and consider a minibatch formed via standard subsampling. We analyze the per-iteration privacy cost under add/remove adjacency; composition and subsampling amplification are handled by the standard RDP accountant and are identical to those used for isotropic DP-SGD.

Let D denote a minibatch dataset and let $g(x) \in \mathbb{R}^k$ be the per-sample gradient of the loss with respect to the equivariant basis coefficients θ for a data point $x \in D$. Let $M = \mathrm{diag}(m_1, \ldots, m_k)$ be the fixed diagonal matrix of generalized He-initialization scales.

For each sample $x \in D$, define the whitened gradient $\tilde{g}(x) = M^{-1} g(x)$, and its clipped version

$$\tilde{g}_{\mathrm{clip}}(x) = \tilde{g}(x) \cdot \min\left(1, \frac{C}{\|\tilde{g}(x)\|_2}\right).$$

By construction, $\|\tilde{g}_{\mathrm{clip}}(x)\|_2 \le C$ for all x.

Define the query function

$$f(D) = \sum_{x \in D} \tilde{g}_{\mathrm{clip}}(x).$$

Under add/remove adjacency, neighboring datasets $D \sim D'$ differ by at most one sample. Therefore,

$$\|f(D) - f(D')\|_2 = \|\tilde{g}_{\mathrm{clip}}(x^\star)\|_2 \le C,$$

where $x^\star$ is the differing sample. The ℓ_2-sensitivity of f is then exactly C.

The whitened mechanism $\mathcal{M}_{\mathrm{white}}$ outputs $\mathcal{M}_{\mathrm{white}}(D) = f(D) + \tilde{Z}$, where $\tilde{Z} \sim \mathcal{N}(0, \sigma^2 C^2 I_k)$. By the Gaussian mechanism, $\mathcal{M}_{\mathrm{white}}$ satisfies Rényi differential privacy of order α with parameter

$$\rho(\alpha) = \frac{\alpha \Delta^2}{2\sigma_G^2},$$

where Δ is the ℓ_2-sensitivity of f and σ_G^2 is the noise variance per coordinate (Mironov, 2017). Substituting $\Delta = C$ and $\sigma_G^2 = \sigma^2 C^2$ yields

$$\rho(\alpha) = \frac{\alpha}{2\sigma^2}.$$

This is the standard RDP parameter for DP-SGD, depending only on the noise multiplier σ and the RDP order α.

The final parameter update $\Delta\theta$ is obtained by applying the deterministic, data-independent transformation $h(u) = M\,u$ to the output of $\mathcal{M}_{\text{white}}$, giving

$$\Delta\theta = h(\mathcal{M}_{\mathrm{white}}(D)).$$

Since h is fixed prior to training and does not depend on the data, the post-processing property of differential privacy implies that the resulting mechanism satisfies the same RDP guarantee as $\mathcal{M}_{\mathrm{white}}$.

Each training iteration therefore incurs RDP parameter $\rho(\alpha) = \alpha/(2\sigma^2)$, identical to that of standard isotropic DP-SGD. Privacy amplification via minibatch subsampling and composition across iterations follow standard RDP accounting and yield the same final (ε, δ) guarantee as in isotropic DP-SGD. □