Implicit Bias of Different Optimizers

Pakeeza Ehsan, Ngwefor Sanda-Ndinyanka, Fabian Gröger
University of Basel

Abstract

The choice of optimization algorithm significantly influences the implicit bias in deep learning models, affecting both convergence properties and generalization performance. Understanding this bias is crucial as different optimizers exhibit distinct tendencies towards specific global minima. This report explores the implicit bias of Unnormalized Sharpness-Aware Minimization (USAM), a variant of Sharpness-Aware Minimization (SAM) that removes gradient normalization, facilitating theoretical analysis. We investigate how USAM selects global minima across two classes of loss functions: convex losses with a unique finite root and strictly monotone losses in linearly separable settings. Our findings reveal that USAM converges to the closest global minimum in terms of L2 distance for convex loss functions, mirroring the behavior of Gradient Descent (GD). In contrast, USAM aligns with the maximum margin solution for strictly monotone losses, akin to GD in classification settings. These insights highlight the role of gradient perturbation in escaping saddle points while maintaining the solution space of the data, contributing to a deeper understanding of optimization-induced biases. Our empirical evaluation on CIFAR-10 corroborates these theoretical findings, demonstrating that while USAM solutions show weaker generalization in binary classification, they occupy flatter minima, potentially contributing to robustness.

1 Introduction

Optimization plays a fundamental role in deep learning by iteratively adjusting model parameters to minimize a given loss function. The choice of optimization algorithm not only determines the rate and direction of convergence but also influences the implicit bias in solution selection, which can impact model generalization. Various optimizers, including Gradient Descent (GD), Stochastic Gradient Descent (SGD), and adaptive methods like Adam and AdaGrad, have been extensively studied in this regard [1–3]. Notably, Sharpness-Aware Minimization (SAM) and its variants, such as Unnormalized Sharpness-Aware Minimization (USAM) and Deterministic Normalization SAM (DNSAM), have been proposed to enhance generalization by seeking flatter minima in the loss landscape [4, 5].

Each optimization algorithm exhibits a distinct implicit bias, favoring specific types of solutions. For instance, GD is known to converge to the minimum L2-norm solution in underdetermined linear regression, while SGD tends to select solutions with simpler structures due to its stochastic updates [1–3]. Understanding such biases is essential for designing robust deep learning models, particularly in high-dimensional, overparameterized settings where multiple global minima exist.

This report focuses on the implicit bias of USAM, a variant of SAM that removes gradient normalization, making it more amenable to theoretical analysis. We analyze how USAM influences solution selection in two key scenarios: convex loss functions with a unique finite root and strictly monotone losses in linearly separable classification problems. Our theoretical findings indicate that USAM behaves similarly to GD, converging to the closest global minimum in convex settings and aligning with the maximum margin separator in classification tasks. These results suggest that USAM's gradient perturbation primarily aids in escaping saddle points without introducing new directions beyond the data manifold.

University of Basel, Course on Foundations of Deep Learning.

To validate our theoretical insights, we conduct empirical experiments on the CIFAR-10 dataset, comparing USAM to GD in binary and multi-class classification settings. Our findings reveal that while USAM solutions exhibit slightly weaker generalization in binary classification, they reside in flatter regions of the loss landscape, a property associated with improved robustness. In multi-class classification, the differences between GD and USAM diminish, suggesting that implicit bias effects become less pronounced in highly underdetermined settings.

2 Related Work

The concept of implicit bias, introduced by optimization methods, was thoroughly investigated by Gunasekar et al. [1]. Their work demonstrated that optimization algorithms inherently prefer specific solutions among multiple global minima, influenced by the geometry of the optimization landscape. They examined gradient descent (GD), mirror descent, and natural gradient descent for linear regression and classification problems. For GD, it was shown that solutions tend to minimize the Euclidean norm in under-determined least squares problems, while for strictly monotone losses, such as logistic loss, GD converges directionally to the maximum margin solution. These findings underscore the pivotal role of algorithmic geometry in guiding convergence, independent of explicit regularization or hyperparameter configurations.

Sharpness-aware minimization (SAM), introduced by Pierre Foret and Neyshabur [4], represents a significant advancement in enhancing generalization performance in deep learning. SAM optimizes for flatter regions in the loss landscape, which are correlated with better generalization. Compagnoni et al. [5] extended this understanding by deriving stochastic differential equations (SDEs) to model SAM and its variants. They revealed that SAM's preference for flat minima arises from Hessian-dependent noise introduced during optimization. Additionally, their work highlighted that SAM, unlike its unnormalized counterpart USAM, is less prone to converging to sharp minima but may exhibit challenges when escaping saddle points. These insights are particularly relevant for understanding the implicit bias of optimizers in non-convex landscapes.

Building on the work by Gunasekar et al. [1], several studies have explored alternative optimization methods to extend theoretical and empirical understanding of implicit bias. For instance, the introduction of unnormalized SAM (USAM) simplified the theoretical analysis of sharpness-aware optimization while retaining essential characteristics. Furthermore, deterministic normalization SAM (DNSAM) was proposed as an intermediate variant to bridge the gap between SAM and USAM, providing a more interpretable framework for theoretical studies.

Neyshabur et al. [3] examined how optimization-induced implicit regularization affects the capacity and generalization of deep neural networks. They argued that factors such as initialization, step size, and the optimization algorithm itself have a more significant impact on generalization than explicit regularization techniques. These findings complement the results of SAM-based studies, emphasizing the importance of exploring implicit mechanisms in modern optimization methods.

3 Theoretical Analysis

In order to theoretically analyze USAM, we replicate the setting and results from Gunasekar et al. [1], specifically for the analysis of gradient descent. We study the bias in terms of two classes of loss functions: losses with a unique finite root and strictly monotone ones.

3.1 Losses with a Unique Finite Root

We first consider learning linear models using losses with a unique finite root, such as the squared loss, where the loss $l(\hat{y}, y)$ between a prediction $\hat{y}$ and label y is minimized at a unique finite value of $\hat{y}$. We assume without loss of generality, that $\min_{\hat{y}} l(\hat{y}, y) = 0$ and the unique minimizer is $\hat{y} = y$.

Property 3.1 (Losses with a unique finite root)**.** For any y a sequence $\{\hat{y}_t\}_{t=1}^{\infty}$ minimizes $l(\cdot, y$, i.e., $l(\hat{y}, y) \xrightarrow{t\to\infty} \inf_{\hat{y}} l(\hat{y}, y) = 0$ if and only if $\hat{y}_t \xrightarrow{t\to\infty} y$.

Denote the training dataset $\{(x_n, y_n) : n = 1, 2, \ldots, N\}$ with features $x_n \in \mathbb{R}^d$ and labels $y_n \in \mathbb{R}$. The empirical loss minimizer of a linear model $f(x) = \langle w, x\rangle$ with parameters $w \in \mathbb{R}^d$ is given by,

$$\min_w \mathcal{L}(w) := \sum_{n=1}^{N} l(\langle w, x_n\rangle, y_n). \tag{1}$$

Like Gunasekar et al. [1], we are interested in the case where $N < d$ and the observations are attainable, i.e., $\min_w \mathcal{L}(w) = 0$. Under these conditions, the optimization problem in eq. 1 is underdetermined and has multiple global minima denoted by $\mathcal{G} = \{w : \mathcal{L}(w) = 0\} = \{w : \forall n, \langle w, x\rangle = y_n\}$.

Consider gradient descent updates for minimizing $\mathcal{L}(w)$ with step-size sequence $\{\eta_t\}_t$ and initialization $w_{(0)}$,

$$w_{(t+1)} = w_{(t)} - \eta_t \nabla \mathcal{L}(w_{(t)}). \tag{2}$$

Gunasekar et al. [1] showed that if $w_{(t)}$ minimizes the empirical loss in eq. 1, then the iterates converge to the unique global minimum that is closest to initialization $w_{(0)}$ in l_2 distance, i.e., $w_{(t)} \rightarrow \arg\min_{w\in\mathcal{G}} ||w - w_{(0)}||_2$. This can be seen as for any w, the gradients $\nabla\mathcal{L}(w) = \sum_n l'(\langle w, x_n\rangle, y_n)x_n$ are always constrained to the fixed subspace spanned by the data $\{x_n\}_n$, and thus the iterates $w_{(t)}$ are confined to the low dimensional affine manifold $w_{(0)} + \text{span}(\{x_n\}_n)$. Within this low dimensional manifold, there is a unique global minimizer w that satisfies the linear constraints in $\mathcal{G} = \{w : \langle w, x_n\rangle = y_n, \forall n \in [N]\}$.

Now consider the USAM updates for minimizing $\mathcal{L}(w)$ with step-size sequence $\{\eta_t\}_t$, initialization $w_{(0)}$, and radius $\rho > 0$,

$$w_{(t+1)} = w_{(t)} - \eta_t \nabla \mathcal{L}\left(w_{(t)} + \rho \nabla \mathcal{L}(w_{(t)})\right). \tag{3}$$

The only difference between the updates for gradient descent (eq. 2) and USAM (eq. 3) lies in the perturbation of the parameters $w_{(t)}$ proportional to the gradient itself. This perturbation step does not introduce new directions outside of the span of the data $\{x_n\}_n$ when assuming $\rho > 0$. As the gradient is evaluated at a point which lies in the span of $w_{(0)} + \text{span}(\{x_n\}_n)$ the resulting updates $w_{(t+1)}$ thus lie within the same fixed subspace where a unique solution exists. This concludes that the iterates of USAM converge to the same unique global minimum as gradient descent, i.e., the one closest to initialization $w_{(0)}$ in l_2 distance. This is illustrated in figure 1 where the dataset $\{(x_1 = [1, 2], y_1 = 1)\}$ is optimized using gradient descent and USAM.

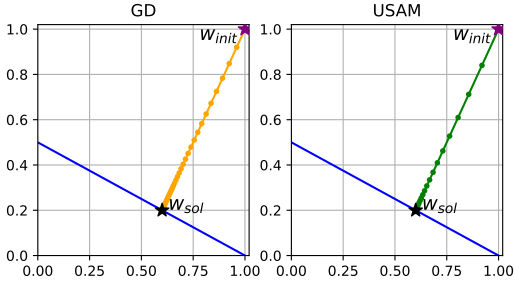

Figure 1: Example showing that gradient descent (left) and USAM (right) converge to the same solution w_{sol} out of the set $\mathcal{G}$ of global minima (blue line) when optimizing $\mathcal{L}(w)$ with dataset $\{(x_1 = [1, 2], y_1 = 1)\}$, squared loss $\ell(u, y) = (u - y)^2$, and initialized with $w_{\text{init}} = [1, 1]$.

3.2 Strictly Monotone Losses

Following the approach of Gunasekar et al. [1], we now shift our focus to strictly monotone loss functions ℓ. In these cases, the behavior of the implicit bias exhibits fundamental differences, as do the conditions under which the implicit bias can be characterized. Such loss functions are common in classification problems where $y = \{-1, 1\}$ and $\ell(f(x), y)$ is a typical continuous surrogate of the 0-1 loss. Examples of such loss functions include logistic loss, exponential loss, and probit loss.

Property 3.2 (Strict monotone losses)**.** $\ell(\hat{y}, y)$ is bounded from below, and $\forall y$, $\ell(\hat{y}, y)$ is strictly monotonically decreasing in $\hat{y}$. Without loss of generality, $\forall y$, $\inf_{\hat{y}} \ell(\hat{y}, y) = 0$ and $\ell(\hat{y}, y) \overset{\hat{y}y\rightarrow\infty}{\longrightarrow} 0$.

We look at classification models that fit the training data $\{x_n, y_n\}_n$ with linear decision boundaries $f(x) = \langle w, x\rangle$ with decision rule given by $\hat{y}(x) = \text{sign}(f(x))$. We also assume without loss of generality that $y_n = 1$ for all n.

We again look at unregularized empirical risk minimization objective of the form in eq. (1), but now with strictly monotone losses with linearly separable data, i.e., $\exists w : \forall n, y_n \langle w, x_n \rangle > 0$, where the empirical loss $\mathcal{L}(w)$ is again ill-posed, and moreover $\mathcal{L}(w)$ does not have any finite minimizer, i.e, $\mathcal{L}(w) \to 0$ only as $\|w\| \to \infty$. Thus, for any sequence $\{w_t\}_{t=0}^{\infty}$, if $\mathcal{L}(w_t) \to 0$, then w_t necessarily diverges to infinity rather than converge, and hence we cannot talk about $\lim_{t\to\infty} w_t$. Instead, we look at the limit direction $\bar{w}_\infty = \lim\limits_{t\to\infty} \frac{w_t}{\|w_t\|}$ whenever the limit exists. We refer to existence of this limit as convergent in direction. Note that the limit direction fully specifies the decision rule of the classifier that we care about.

Soudry et al. [2] showed that for almost all linearly separable datasets, gradient descent with *any initialization and any bounded step-size* converges in direction to maximum margin separator with unit ℓ_2 norm, i.e., the hard margin support vector machine classifier,

$$\bar{w}_\infty = \lim_{t\to\infty} \frac{w_t}{\|w_t\|_2} = w^*_{\|\cdot\|_2} := \underset{\|w_t\|_2 \le 1}{\arg\max} \min_n y_n \langle w, x_n \rangle . \tag{4}$$

For standard gradient descent the gradient in this setting is $\nabla\mathcal{L}(w_{(t)}) = \sum_n \ell'(y_n w_{(t)}^\top x_n) y_n x_n = \sum_n \ell'(w_{(t)}^\top x_n) x_n$ and is therefore a weighted sum of the data points $\{x_n\}_n$. Thus $\nabla\mathcal{L}(w_{(t)})$ always lies in the span of the dataset, i.e., $\text{span}(\nabla\mathcal{L}(w_{(t)})) = \text{span}(\{x_n\}_n)$. The updates for USAM, given in eq. (3), can be stated using the gradient descent updates from eq. (2) at a perturbed point $\tilde{w}_{(t)} = w_{(t)} + \rho\nabla\mathcal{L}(w_{(t)})$. The gradient evaluated at this perturbed point is $\nabla\mathcal{L}(\tilde{w}_{(t)}) = \sum_n \ell'(y_n \tilde{w}_{(t)}^\top x_n) y_n x_n = \sum_n \ell'(\tilde{w}_{(t)}^\top x_n) x_n$. So with this perturbed point, the gradient lies in the span of the dataset, i.e., $\text{span}(\nabla\mathcal{L}(\tilde{w}_{(t)})) \subseteq \text{span}(\nabla\mathcal{L}(w_{(t)})) = \text{span}(\{x_n\}_n)$. In gradient descent, Soudry et al. [2] showed that the direction $\bar{w}_\infty$ asymptotically aligns with the maximum margin solution $w^*_{\|\cdot\|_2}$ (eq. 4) and for USAM, $\nabla\mathcal{L}(\tilde{w}_{(t)})$ lies in the same span as $\nabla\mathcal{L}(w_{(t)})$ and the asymptotic contributions are still dominated by the same support vectors. The perturbation introduced by ρ only scales the updates but does not change their asymptotic direction. In conclusion, USAM also converges in the direction of the maximum margin separator in this setting.

4 Empirical Analysis

Table 1: Results of the empirical analysis when comparing the solutions of a linear model trained on flattened CIFAR-10 images obtained through GD and USAM measured in terms of generalization gap ($\Delta\mathcal{L}$), trace of the hessian ($\text{tr}[\nabla^2\mathcal{L}]$), perturbation sensitivity ($|\nabla\mathcal{L}|$) in terms of mean μ and standard deviation σ, norm of the solution ($\|\mathbf{w}_{\text{sol}}\|$), and distance to the initial solution ($\|\mathbf{w}_{\text{init}} - \mathbf{w}_{\text{sol}}\|$). Lower is better for all reported metrics except for the distance to the initial solution.

Optimizer	$\Delta\mathcal{L}$	$\text{tr}[\nabla^2\mathcal{L}]$	$\lvert\nabla\mathcal{L}\rvert$	$\lVert\mathbf{w}_{\text{sol}}\rVert$	$\lVert\mathbf{w}_{\text{init}} - \mathbf{w}_{\text{sol}}\rVert$
Binary classification					
GD ($\mathbf{w}_{\text{init}} = \mathbf{0}, \eta = 0.1$)	1.13	12.49	$2.51 \times 10^{-9}, 5.93 \times 10^{-9}$	10.47	10.47
GD ($\mathbf{w}_{\text{init}}$ rand, $\eta = 0.1$)	1.14	12.48	$1.06 \times 10^{-9}, 5.52 \times 10^{-9}$	10.68	10.41
USAM ($\mathbf{w}_{\text{init}} = \mathbf{0}, \eta = 0.1, \rho = 0.1$)	1.14	12.42	$1.03 \times 10^{-10}, 5.84 \times 10^{-9}$	10.52	10.53
USAM ($\mathbf{w}_{\text{init}}$ rand, $\eta = 0.1, \rho = 0.1$)	1.15	12.40	$1.67 \times 10^{-10}, 5.72 \times 10^{-9}$	10.74	10.47
USAM ($\mathbf{w}_{\text{init}} = \mathbf{0}, \eta = 0.1, \rho = 0.01$)	1.14	12.49	$3.23 \times 10^{-10}, 5.79 \times 10^{-9}$	10.47	10.48
USAM ($\mathbf{w}_{\text{init}}$ rand, $\eta = 0.1, \rho = 0.01$)	1.14	12.47	$1.99 \times 10^{-10}, 6.24 \times 10^{-9}$	10.67	10.47
Multi-class classification					
GD ($\mathbf{w}_{\text{init}} = \mathbf{0}, \eta = 0.1$)	1.14	164.39	$3.16 \times 10^{-9}, 7.22 \times 10^{-9}$	36.09	36.10
GD ($\mathbf{w}_{\text{init}}$ rand, $\eta = 0.1$)	1.14	164.31	$1.73 \times 10^{-9}, 8.80 \times 10^{-9}$	37.03	36.05
USAM ($\mathbf{w}_{\text{init}} = \mathbf{0}, \eta = 0.1, \rho = 0.1$)	1.14	164.34	$4.99 \times 10^{-10}, 8.51 \times 10^{-9}$	36.14	36.16
USAM ($\mathbf{w}_{\text{init}}$ rand, $\eta = 0.1, \rho = 0.1$)	1.14	164.26	$5.88 \times 10^{-10}, 8.99 \times 10^{-9}$	37.09	36.11
USAM ($\mathbf{w}_{\text{init}} = \mathbf{0}, \eta = 0.1, \rho = 0.01$)	1.14	164.39	$5.04 \times 10^{-10}, 8.62 \times 10^{-9}$	36.08	36.11
USAM ($\mathbf{w}_{\text{init}}$ rand, $\eta = 0.1, \rho = 0.01$)	1.14	164.32	$2.04 \times 10^{-10}, 8.13 \times 10^{-9}$	36.96	36.12

We conduct an empirical investigation to assess the bias of USAM based on the previously mentioned loss functions and compare the results with that of GD. The experimental settings are adjusted to align with the assumptions made in the theoretical analysis. For this investigation, we rely on the CIFAR-10 dataset and train a linear model $f_{\mathbf{w}}(\cdot) = (\cdot)\mathbf{w}$, where $\mathbf{w} \in \mathbb{R}^{d\times c}$ with feature dimensions d and number of classes c, using cross-entropy loss directly on the flattened grayscale images, i.e.,

$d = 32 \times 32 = 1024$. We sample 100 points per class, resulting in $N = 1000$, giving us the setting we theoretically analyzed, namely $N < d$. The same subsampling is repeated for the test set, which gives a 50/50 split. Furthermore, we consider two settings: one in which we limit the dataset to only two classes, specifically "plane" and "car," and another where no such restriction is applied. . This enables us to investigate the influence of switching from a binary classification scenario to a multi-class setting.

We conduct the experiment for a linear model initialized at zero ($\mathbf{w}_{\text{init}} = \mathbf{0}$) and at random ($\mathbf{w}_{\text{init}}$ rand). The model is trained using both GD and USAM for 500 iterations, using fixed learning rate $\eta = 0.1$ and varying perturbation amounts $\rho \in \{0.01, 0.1\}$ Results are measured in terms of generalization gap (i.e., difference in loss for the final solution on the training and test set samples), trace of the hessian, perturbation sensitivity (i.e., computed by adding small random noise (1×10^{-5}) of fixed Frobenius norm to the weights, and measuring how much the loss increases on average for 100 trails), L2 norm of the final solution ($\|\mathbf{w}_{\text{sol}}\|$), and distance to the initial solution measured in terms of L2 norm ($\|\mathbf{w}_{\text{init}} - \mathbf{w}_{\text{sol}}\|$).

Table 1 shows the results of the empirical analysis for both binary and multi-class classification. In the binary classification setting, the norm of the solution and its distance from initialization differs significantly between GD and USAM, with USAM exhibiting larger values. Additionally, USAM shows a wider generalization gap compared to GD. However, when examining the trace of the hessian and perturbation sensitivity, we observe that while USAM generates weaker solutions compared to GD in this scenario, the solutions reside in flatter minima, as the trace of the hessian is slightly lower and perturbations have a smaller impact on the loss. This difference in generalization performance may stem from the fact that USAM, by introducing perturbations, navigates toward flatter minima that may not necessarily align with optimal generalization in this restricted setting. Interestingly, the generalization gap for both optimizers is identical in the multi-class setting, and the overall differences between their solutions are minimal. The only exception is the flatness of the solution, where USAM has a slightly lower trace of the hessian and a perturbation sensitivity an order of magnitude smaller than GD. This suggests that examining an even more under-determined scenario, such as multi-class classification, aligns more closely with the theoretical analysis, as both optimizers produce similar solutions. This is confirmed by comparing the L2 norm of the solutions from both optimizers, where $3.02 \times 10^{-5} \pm 1.73 \times 10^{-7}$ indicates that the solutions are indeed very close.

5 Conclusion

In this report, we examined the implicit bias of the Unnormalized Sharpness-Aware Minimization (USAM) optimization algorithm. Our theoretical analysis demonstrated that USAM, like Gradient Descent (GD), converges to the unique global minimum closest to initialization for convex loss functions and aligns with the maximum margin separator for strictly monotone losses. The empirical analysis showed that in a binary classification setting, the differences between the solutions obtained through USAM and GD are vastly different, whereas for multi-class classification, the difference is minimal, and they converge to solutions that are very close to each other. Despite these sometimes minimal differences, we observed that USAM consistently produces solutions in flatter regions.

In future work, it would be interesting to repeat the experiment, but instead of using the flattened images as features, one uses extracted embeddings from a pre-trained model, such as one trained on ImageNet. Even if the analyzed setting here is very different from real-world use cases, this experiment would, on the other hand, be similar to what is currently being done in transfer learning and representation learning. Moreover, since this will more likely yield linearly separable samples, it will align with the assumptions of the theoretical analysis even more.

Limitations

Note that in the theoretical analysis we assume each update uses a sufficiently small step-size for both gradient descent and USAM and for both losses with a unique finite root and strictly monotone ones. Additionally, for USAM, we assume that the perturbation parameter ρ is chosen sufficiently small so that it does not alter the overall direction of the gradient and cause erratic jumps.

We acknowledge that the setting in the empirical analysis is idealized and vastly differs from real use cases, as only a single layer is trained on flattened images here. However, we are still under the impression that this idealized setting sheds light on some insights into the optimization behavior of the different optimizers.

References

[1] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing Implicit Bias in Terms of Optimization Geometry. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1832–1841. PMLR, July 2018. URL `https://proceedings.mlr.press/v80/gunasekar18a.html`. ISSN: 2640-3498.

[2] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro. The Implicit Bias of Gradient Descent on Separable Data. February 2018. URL `https://openreview.net/forum?id=r1q7n9gAb`.

[3] Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of Optimization and Implicit Regularization in Deep Learning, May 2017. URL `http://arxiv.org/abs/1705.03071`. arXiv:1705.03071 [cs].

[4] Hossein Mobahi Pierre Foret, Ariel Kleiner and Behnam Neyshabur. Sharpness-Aware Minimization for Efficiently Improving Generalization, 2021. URL `https://arxiv.org/abs/2010.01412`. arXiv:2010.01412v3 [cs.LG].

[5] Enea Monzio Compagnoni, Luca Biggio, Antonio Orvieto, Frank Norbert Proske, Hans Kersting, and Aurelien Lucchi. An SDE for Modeling SAM: Theory and insights, 2023. URL `https://proceedings.mlr.press/v202/monzio-compagnoni23a/monzio-compagnoni23a.pdf`. arXiv:2301.08203v3 [cs.LG].