OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Rethinking Sharpness-Aware Minimization as Variational Inference

Szilvia Ujváry [1,2,*] SRU23@CAM.AC.UK
Zsigmond Telek [2,*] ZT820@IC.AC.UK
Anna Kerekes [1,*] AK2229@CAM.AC.UK
Anna Mészáros [1,*] AM3049@CAM.AC.UK
Ferenc Huszár [1] FH277@CAM.AC.UK
[1]*University of Cambridge, U.K.;* [2]*Imperial College London, U.K.;* [*]*equal contributions*

Abstract

Sharpness-aware minimisation (SAM) aims to improve the generalisation of gradient-based learning by seeking out flat minima. In this work, we establish connections between SAM and mean-field variational inference (MFVI) of neural network parameters. We show that both these methods have interpretations as optimizing notions of flatness, and when using the reparametrisation trick, they both boil down to calculating the gradient at a perturbed version of the current mean parameter. This thinking motivates our study of algorithms that combine or interpolate between SAM and MFVI. We evaluate the proposed variational algorithms on several benchmark datasets, and compare their performance to variants of SAM. Taking a broader perspective, our work suggests that SAM-like updates can be used as a drop-in replacement for the reparametrisation trick.

1. Introduction

The training loss of overparametrized models often has several equally low minima, which may differ widely in their performance at test time. The most salient effort to characterise which minima lead to better generalisation is the one that links *flatness* to generalisation [3–5, 9–11]. Motivated by this, SAM aims to directly bias optimization towards flat regions leading to substantial improvement in generalisation while only doubling the computational cost per iteration.

Other methods, such as MFVI and Evolution Strategies (ES) also have *flatness-seeking* interpretations. The pursuit of flat minima was originally motivated by a Bayesian, minimum description length (MDL) argument [9]. Thus, it is not surprising that minimizing a variational Bayesian objective, itself an estimate of the description length of training data [8], favours flat minima. This connection between variational Bayesian deep learning and SAM has remained largely unexplored. In this work we show that the connections between MFVI and SAM are more than philosophical.

Concrete implementations of SAM and MFVI (with the reparametrization trick) both calculate gradients at a perturbed point around the current parameter. While SAM takes the worst-case perturbation, MFVI uses a random Gaussian perturbation. In this sense, the SAM update can be thought of as a biased deterministic approximation replacing the unbiased but high-variance single-sample Monte Carlo gradient in MFVI. Exploring this further, we make the following contributions:

1. We establish connections between SAM and MFVI, characterising and comparing their flatness-seeking inductive biases (Appendix A, B).
2. We propose the algorithm VariationalSAM, which combines aspects of SAM and MFVI (C).
3. We compare all methods on benchmark datasets (CIFAR-10 and CIFAR-100).

2. Sharpness-Aware Minimization and Mean-Field Variational Inference

2.1. Sharpness-Aware Minimization (SAM)

In this section, we present SAM with a slight change in notation relative to original works in order to make connections to variational methods more apparent later. Denoting the parameter by μ, and the number of parameters by p, the idealized loss function SAM aims to minimize can be written as

$$L_{\text{SAM}}(\mu, \Sigma) = L(\mu) + \max_{\epsilon^T\Sigma^{-1}\epsilon\leq p} [L(\mu+\epsilon) - L(\mu)] + \alpha\|\mu\|_2^2 \tag{1}$$

Σ is any positive definite matrix. Setting $\Sigma = \frac{\rho^2}{p}I$, where ρ is a constant parameter recovers the standard SAM [6], while setting Σ as a function of μ, recovers newer variants e.g. Adaptive SAM (ASAM) [15] and Fisher SAM (FSAM) [12]. The parameter $\alpha > 0$ controls L_2 regularization. Computing gradients of (1) with respect to μ is intractable in general, hence an approximation is used. Taking a first-order Taylor expansion of L we can approximate the optimal ϵ [6, 12, 15]:

$$\epsilon^*(\mu) = \sqrt{p}\frac{\Sigma\nabla_\theta^\top L(\mu)}{\sqrt{\nabla_\theta^\top L(\mu)\Sigma\nabla_\theta L(\mu)}} \tag{2}$$

where $\nabla_\theta^\top L(\mu)$ denotes the gradient of the loss evaluated at μ. For reasons that will be clear later, we introduce new notation $\eta(\mu)$ and re-express ϵ^* from η as follows:

$$\epsilon^*(\mu) = \Sigma^{\frac{1}{2}}\eta(\mu), \text{ where } \eta(\mu) = \frac{\sqrt{p}\tilde{g}(\mu)}{\|\tilde{g}(\mu)\|_2} \text{ and } \tilde{g}(\mu) = \Sigma^{\frac{1}{2}}\nabla_\theta^\top L(\mu) \tag{3}$$

The SAM update direction is obtained by plugging this estimate back into (1) and ignoring the dependence of η on μ when differentiating the resulting expression:

$$\nabla_\mu^\top L_{\text{SAM}} \approx \nabla_\theta^\top L(\mu + \Sigma^{\frac{1}{2}}\eta(\mu)) + 2\alpha\mu \tag{4}$$

Note that computing this update direction requires backpropagation twice - first to calculate $\eta(\mu)$, and second to evaluate $\nabla_\theta^\top L$ - thus doubling the computational cost.

2.2. Mean-Field Variational Inference (MFVI)

We can also take an approximate Bayesian approach to inferring model parameters from data. We start with a prior distribution, $p(\theta)$ over the parameters θ, which we choose to be $\mathcal{N}(0, \sigma_0 I)$. We then attempt to infer the posterior distribution of weights using the Bayes' rule, $p(\theta|\mathcal{D}) \propto e^{-NL(\theta)}p(\theta)$, where N denotes the number of data points and we assumed that the loss function L calculates the average negative log likelihood of i. i .d. observations. This operation being intractable, we use an approximate posterior $q(\theta)$, chosen to be Gaussian with mean μ and covariance Σ.

Minimizing the KL-divergence between this approximate q and the true posterior yields the following objective function for μ and Σ:

$$L_{\text{MFVI}}(\mu, \Sigma) = \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\Sigma)}L(\theta) + \frac{1}{N}\text{KL}\left[\mathcal{N}(\mu, \Sigma)\|\mathcal{N}(0, \sigma_0^2 I)\right] \tag{5}$$

When Σ is chosen to be diagonal, the minimization of (5) is called mean-field variational inference (MFVI). Here, we liberally extend this name to algorithms where Σ is non-diagonal. Dropping the KL term from (5) yields variational optimization (VO) [22] and evolution strategies [7, 19, 23].

We can estimate the gradients of L_{MFVI} using the reparametrization trick [13]. We also expand the KL divergence term, and substitute $\alpha := \frac{1}{N\sigma_0^2}$ to obtain:

$$\nabla_\mu^\top L_{\text{MFVI}}(\mu, \Sigma) = \mathbb{E}_{\eta\sim\mathcal{N}(0,I)} \nabla_\theta^\top L(\mu + \Sigma^{\frac{1}{2}}\eta) + 2\alpha\mu. \tag{6}$$

We approximate this by a single-sample Monte Carlo estimate and get the update direction:

$$\nabla_\mu^\top L_{\text{MFVI}}(\mu, \Sigma) \approx \nabla_\theta^\top L(\mu + \Sigma^{\frac{1}{2}}\eta) + 2\alpha\mu, \text{ where } \eta \sim \mathcal{N}(0, I) \tag{7}$$

Note the similarity between Eqns. (4) and (7). They are of precisely the same form except for the perturbation η: in SAM, η is chosen to be the worst-case perturbation calculated deterministically from μ by taking a gradient, while in MFVI, η is drawn from a standard normal. Throughout this work, we will refer to the variant of MFVI where we fix $\Sigma = \frac{\rho^2}{d}I$ as *RandomSAM*, or $L_{\text{RSAM}}(\mu)$.

2.3. Exploring the relationship

Name	perturbation	covariance	$\Sigma =$	Penalty
SAM [6]	worst-case	fixed	$\frac{\rho^2}{p}I$	$\text{KL} = L_2$
Random SAM (MFVI μ only)	Gaussian	fixed	$\frac{\rho^2}{p}I$	KL
MFVI	Gaussian	learned	$\text{diag}(\sigma_i)$	KL
Variational SAM	worst-case	learned	$\text{diag}(\sigma_i)$	KL
Adaptive SAM [15]	worst-case	μ-adaptive	$\text{diag}(\lvert\frac{1}{\mu_i}\rvert)$	L_2
Fisher SAM [12]	worst-case	μ-adaptive	$\text{diag}(F(\mu))$	L_2
Evolution Strategy (ES) [2, 19]	Gaussian	fixed	$\frac{\rho^2}{p}I$	none
CMA-ES [7], VO [22], NES [23]	Gaussian	learned	$\text{diag}(\sigma_i)$	none

Table 1: An overview of methods mentioned in this work. All methods can be related to MFVI or SAM by changing the perturbation type, shape of Σ, or penalty terms. Random SAM is a variation of MFVI with fixed $\Sigma = \frac{\rho^2}{p}I$ and Variational SAM is the version of MFVI, where we learn Σ.

SAM as an upper bound on the Variational Objective: The similarity between the SAM and MFVI updates is not surprising given that L_{SAM} can be considered as a loose upper bound on L_{MFVI} when the variance Σ is sufficiently small, and the number of parameters, p is sufficiently large. This is because in high dimensions, samples from $\mathcal{N}(\mu, \Sigma)$ concentrate around the ellipsoid $(x - \mu)^\top\Sigma^{-1}(x - \mu) = p$. Thus, any expectation over the $\mathcal{N}(\mu, \Sigma)$ can be upper bounded by the maximum value within the ellipsoid i. e. $\max_{(x-\mu)^\top\Sigma^{-1}(x-\mu)\leq p} L(\theta)$

This upper bound relationship is exploited in the proof of theorems used to justify SAM[6], ASAM[15] and FSAM [12]. Those proofs therefore imply an even stronger theoretical justification for MFVI as a means to achieve good generalisation.

Implicit regularization As discussed, both SAM and MFVI have flatness-seeking inductive biases. But can we compare these biases, or connect them to notions of sharpness we understand

	SAM	MFVI
Analysis of idealized loss (what it says on the box)	$L(\mu) + \sqrt{p}\|\|g(\mu)\|\|_\Sigma$	$L(\mu) + \frac{1}{2}\text{Tr}[\Sigma H(\mu)]$
Analysis SGD dynamics (what the algorithm does)	$L(\mu) + \sqrt{p}\|\|g(\mu)\|\|_\Sigma + \frac{\delta}{4}\|\|g(\mu)\|\|_2^2$	$L(\mu) + \frac{\delta}{4}\text{Tr}[\Sigma H(\mu)^2] + \frac{\delta}{4}\|\|g(\mu)\|\|_2^2$

Table 2: Summary of the flatness-seeking regularization properties of SAM and MFVI.

better? Here, we characterize the implicit regularization towards flat minima in terms of higher order derivatives, borrowing techniques from [1, 20, 21]. We summarise these in the following four propositions (we set $\alpha = 0$ ignoring the L_2 or KL penalties for readability): In the **first row** we approximate the idealized loss functions of SAM (Eqn. (1)) and MFVI (Eqn. (5)) in terms of higher order derivatives. SAM approximately penalizes the norm of the gradient, while MFVI penalizes the trace of the Hessian. Near minima, Tr$[\Sigma H(\mu)]$ is a good measure of sharpness, but it can take negative values around saddle points and local maxima. In the **second row** we approximate the implicit regularization properties of stochastic gradient descent with SAM (Eqn. (4)) or MFVI gradients (Eqn. (7)), with small but finite learning rate δ. The $\frac{\delta}{4}\|g(\mu)\|$ term represents the implicit bias of GD with finite step size as in as in [21]. SAM's flatness penalty is the same as the idealized, but for MFVI the penalty contains the trace of the Hessian squared, which is now always positive. These results show us the regularizing properties of both SAM and MFVI. The detailed theorems and proof can be found in Appendix A, B. In Figure 1 we illustrate the sharpness-avoidance of L_{SAM} and L_{RSAM} on a 2D toy example originally introduced in [12]. We can see that both transformations are effective at reducing the attraction basin of the sharp minimum. However, it is still a question which translates better to high-dimensional problems we encounter in deep learning, and which one leads to better generalization.

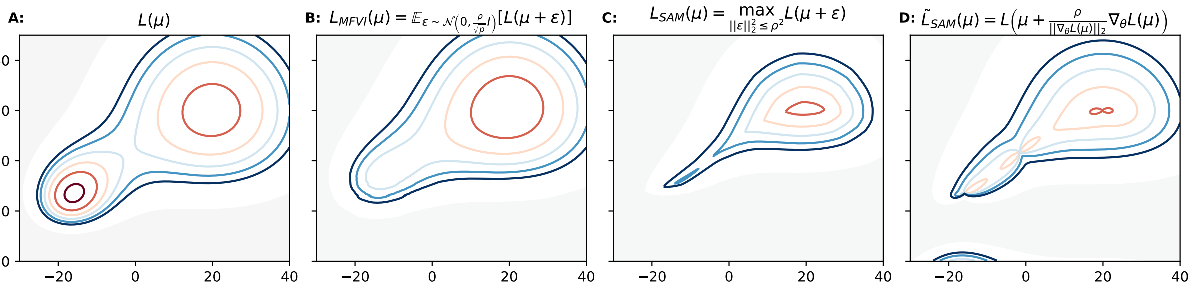

Figure 1: Illustration of MFVI's and SAM's bias towards flatter minima on a 2D toy example. **A**: The original loss L [12], has two minima, a sharp and a wide one. **B:** in MFVI, averaging over the Gaussian variational posterior smoothes out the sharper minimum, increasing the attraction basin of the flat one. **C:** considering the worst-case within a Euclidean ball achieves a similar transformation **D:** the SAM update relies on a Taylor approximation, which does not apply in highly non-linear regions. (in all cases $\rho = 8$).

The above connections between MFVI and SAM motivate our two main research questions which we investigate in the rest of this report:

1. All things being equal, is SAM or `RandomSAM` (MFVI with fixed $\Sigma = \frac{\rho^2}{p}I$ more effective at finding minima that generalise well?

2. It is possible to learn Σ in SAM via gradient descent along with μ as in MFVI. This results in an algorithm we call `VariationalSAM`. Does this have advantages over SAM where Σ is fixed. (Details of the algorithm are in Appendix C, E with a PAC-Bayes.)

3. Image Classification Experiments

In this section, we empirically assess the generalisation performance of the mentioned algorithms: vanilla SGD, SAM, RandomSAM (RSAM) and VariationalSAM (VSAM). Our implementations of these methods is open source, and can be accessed in this repository. Details of the parameters of the experiments can be found in Appendix D.

	CIFAR-10		CIFAR-100	
	WideResNet 28-2	WideResNet 28-10	WideResNet 28-2	WideResNet 28-10
SGD	$95.90^{\pm 0.07}$	$96.97^{\pm 0.12}$	$74.32^{\pm 0.12}$	$80.23^{\pm 0.0040}$
SAM	$96.10^{\pm 0.11}$	$97.20^{\pm 0.07}$	$76.25^{\pm 0.27}$	$83.26^{\pm 0.0004}$
VSAM	$94.18^{\pm 0.11}$	$96.95^{\pm 0.42}$	$74.74^{\pm 0.29}$	$81.72^{\pm 0.0200}$
RSAM	$96.08^{\pm 0.13}$	$97.11^{\pm 0.05}$	$75.78^{\pm 0.05}$	$80.58^{\pm 0.0006}$

Table 3: Image Classification

SAM visibly outperforms all other methods in all experiments. However, on CIFAR-10, RandomSAM performs almost as well as SAM. Given that RandomSAM is a much simpler method operating with random noise and only one backpropagation, this result somewhat weakens the advantages of SAM. It raises the question whether the superior results of SAM on larger datasets could be reproduced by other methods using random perturbations in the parameter space. On CIFAR-100 we see a significant gap in performance, however this may be due to our limited computing budget not allowing for more extensive grid search of parameters.

Our adaptive version of SAM, Variational SAM, does not appear to work better than SAM. This may be because during training, Σ starts to increase, and often reaches a magnitude where the Taylor approximation underlying SAM no longer holds. This raises the question whether random MFVI would have an advantage over SAM when more flexible covariance structures are used.

Although the evaluation of Adaptive SAM and Fisher SAM are not in the focus of this paper, we have included a limited set of results in Appendix G for comparison.

4. Summary and Future Work

In this work, we have provided a novel interpretation of SAM, from the angle of Variational Inference. This led to the comparison of SAM with methods taken from Variational Inference: RandomSAM (Variational Optimization) and Variational SAM. The latter method performed rather unpromisingly. Interestingly, RandomSAM performed similarly to SAM on the CIFAR-10 dataset. This raises questions about optimality of SAM in biasing the optimisation against sharp minima, which we would like to explore in the future. For example, one could construct an algorithm that interpolates between SAM and RandomSAM by adding random noise to the SAM perturbation.

In most experiments, a limitation of computational power has prevented extensive grid-searches on the hyperparameters in our models. Therefore, the reported test accuracies can likely be improved with further parameter calibration.

References

[1] David G. T. Barrett and Benoit Dherin. Implicit gradient regularization. *CoRR*, abs/2009.11162, 2020. URL https://arxiv.org/abs/2009.11162.

[2] H.G. Beyer and HP. Schwefel. Evolution strategies - a comprehensive introduction. *Natural Computing*, 1(1):3–52, March 2002. doi: 10.1023/A:1015059928466.

[3] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer T. Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *CoRR*, abs/1611.01838, 2016. URL http://arxiv.org/abs/1611.01838.

[4] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *CoRR*, abs/1703.04933, 2017. URL http://arxiv.org/abs/1703.04933.

[5] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *CoRR*, abs/1703.11008, 2017. URL https://arxiv.org/abs/1703.11008.

[6] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *CoRR*, abs/2010.01412, 2020. URL https://arxiv.org/abs/2010.01412.

[7] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001. doi: 10.1162/106365601750190398.

[8] Marton Havasi, Robert Peharz, and José Miguel Hernández-Lobato. Minimal random code learning: Getting bits back from compressed model parameters, 2018. URL https://arxiv.org/abs/1810.00440.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Comput.*, 9(1):1–42, jan 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.1. URL https://doi.org/10.1162/neco.1997.9.1.1.

[10] Yiding Jiang, Parth Natekar, Manik Sharma, Sumukh K. Aithal, Dhruva Kashyap, Natarajan Subramanyam, Carlos Lassance, Daniel M. Roy, Gintare Karolina Dziugaite, Suriya Gunasekar, Isabelle Guyon, Pierre Foret, Scott Yak i, Hossein Mobahi, Behnam Neyshabur, and Samy Bengio. Methods and analysis of the first competition in predicting generalization of deep learning. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, 2021. URL http://proceedings.mlr.press/v133/jiang21a.html.

[11] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2016. URL http://arxiv.org/abs/1609.04836.

[12] Minyoung Kim, Da Li, Shell Xu Hu, and Timothy M. Hospedales. Fisher sam: Information geometry and sharpness aware minimisation, 2022. URL https://arxiv.org/abs/2206.04920.

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

[14] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[15] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. *CoRR*, abs/2102.11600, 2021. URL https://arxiv.org/abs/2102.11600.

[16] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. ISSN 00905364. URL http://www.jstor.org/stable/2674095.

[17] David A. McAllester. Pac-bayesian model averaging. In *COLT '99*, 1999.

[18] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[19] I. Rechenberg. Evolutionsstrategien. In Berthold Schneider and Ulrich Ranft, editors, *Simulationsmethoden in der Medizin und Biologie*, pages 83–114, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg. ISBN 978-3-642-81283-5.

[20] Daniel A Roberts. Sgd implicitly regularizes the generalization error. In *Workshop on Integration of Deep Learning Theories at NeurIPS 2018*, 2018.

[21] Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=rq_Qr0c1Hyo.

[22] Joe Staines and David Barber. Variational optimization, 2012. URL https://arxiv.org/abs/1212.4507.

[23] Daan Wierstra, Tom Schaul, Jan Peters, and Juergen Schmidhuber. Natural evolution strategies. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, pages 3381–3387, 2008. doi: 10.1109/CEC.2008.4631255.

[24] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Edwin R. Hancock Richard C. Wilson and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, September 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87. URL https://dx.doi.org/10.5244/C.30.87.

5. Appendix

A. Implicit regularization of SAM

Proposition 1 *The following approximation holds for the SAM objective ($\Sigma = \frac{\rho^2}{p} I$)*

$$L_{SAM}(\mu) \approx L(\mu) + \rho \|\nabla_\theta L(\mu)\|_2. \tag{8}$$

Furthermore, for general Σ, this takes the form

$$L_{SAM}(\mu, \Sigma) \approx L(\mu) + \sqrt{p}\|\nabla_\theta L(\mu)\|_\Sigma = L(\mu) + \sqrt{p}\sqrt{\nabla_\theta^\top L(\mu)\Sigma\nabla_\theta L(\mu)}. \tag{9}$$

Proof Plugging the solution ϵ^* in (Eqn. (2)) into the SAM objective and performing a first-order Taylor approximation, we arrive at

$$\begin{aligned} L_{\text{SAM}}(\mu, \Sigma) &= \max_{\epsilon^T \Sigma^{-1} \epsilon < p} L(\mu + \epsilon) \\ &= L(\mu + \epsilon^*) \\ &= L(\mu) + \epsilon^{*\top} \nabla_\theta L(\mu) + O(\epsilon^{*\top}\epsilon^*) \\ &= L(\mu) + \sqrt{p} \frac{\nabla_\theta^\top L(\mu)\Sigma}{\sqrt{\nabla_\theta^\top L(\mu)\Sigma\nabla_\theta L(\mu)}} \nabla_\theta L(\mu) + O(\epsilon^{*\top}\epsilon^*) \\ &= L(\mu) + \sqrt{p}\sqrt{\nabla_\theta^\top L(\mu)\Sigma\nabla_\theta L(\mu)} + O(\epsilon^{*\top}\epsilon^*), \end{aligned}$$

where we have used that Σ is a symmetric matrix. In order to show that the error term is controlled, let us use express ϵ^* as in (Eqn. (3)) as $\epsilon^* = \Sigma^{\frac{1}{2}}\eta$. Then

$$\epsilon^{*\top}\epsilon^* = \eta^\top \Sigma \eta. \tag{10}$$

From (Eqn. (3)), we see that $\|\eta\|_2 = \sqrt{p}$, which means that as long as the largest (in magnitude) eigenvalue of Σ, $\lambda_{\text{max}} < K\frac{1}{p}$ for some K constant, we have $\eta^\top \Sigma \eta < K$. In the special case of SAM, we have $\Sigma = \frac{\rho^2}{p} I$, hence the error term scales with ρ^2. ■

Proposition 2 *Full batch gradient descent using the SAM step follows a path that is closest to the exact continuous path given by $\dot{\mu} = -\nabla_\mu \tilde{L}_{SAM}(\mu)$, where $\tilde{L}_{SAM}(\mu)$ is given by*

$$\tilde{L}_{SAM}(\mu) \approx L(\mu) + \sqrt{p}\|\nabla_\theta L(\mu)\|_\Sigma + \frac{\delta}{4}||\nabla_\theta L(\mu)||_2^2, \tag{11}$$

where δ is the stepsize of the gradient descent algorithm.

Proof In the proof, we follow [1, 21] in finding a modified loss surface, along which the exact path of gradient flow is closer to the discrete steps of gradient descent on the approximated gradient of the SAM objective (Eqn. (4)). The general formula in [1, 21] of the modified loss surface for loss function $E(\mu) := L_{\text{SAM}}(\mu)$ is

$$\tilde{L}_{\text{SAM}}(\mu) = L_{\text{SAM}}(\mu) + \frac{\delta}{4}||\nabla_\theta L_{\text{SAM}}(\mu)||_2^2. \tag{12}$$

Now we can use the approximation in Proposition 1. In the following, we concentrate on the general case (general Σ). For readability, we use the notation $\nabla_\theta L_{\text{SAM}}(\mu) = g(\mu)$.

$$\tilde{L}_{\text{SAM}}(\mu) = L(\mu) + \sqrt{p}\|g(\mu)\|_2 + \frac{\delta}{4}\Big|\Big|g(\mu) + \sqrt{p}\nabla_\theta\|g(\mu)\|_\Sigma\Big|\Big|_2^2. \tag{13}$$

We can approximate the rightmost term as

$$\Big|\Big|g(\mu) + \sqrt{p}\nabla_\theta\|g(\mu)\|_\Sigma\Big|\Big|_2^2 = \big(g^\top(\mu) + \sqrt{p}\nabla_\theta^\top\|g(\mu)\|_\Sigma\big)\big(g(\mu) + \sqrt{p}\nabla_\theta\|g(\mu)\|_\Sigma\big) \approx \|g(\mu)\|_2^2, \tag{14}$$

since the first two of the remaining terms scale with $\sqrt{p}\Sigma^{\frac{1}{2}}$, which can be made smaller than ρ for sufficiently small Σ. The last remaining term scales with $p\Sigma$, which can be made smaller than ρ^2. Since we also have the scaling factor δ in (Eqn. (13)), we may neglect these terms. ■

B. Sharpness avoidance of variational optimization

Proposition 3 *The following approximation holds for the RandomSAM with constant $\Sigma = \frac{\rho^2}{p}I$:*

$$L_{RSAM}(\mu) \approx L(\mu) + \frac{\rho^2}{2p}TrH(\mu) \tag{15}$$

where $H(\mu)$ denotes the Hessian. For MFVI with general (but symmetric positive definite) Σ we have the following approximation:

$$L_{MFVI}(\mu, \Sigma) \approx L(\mu) + \frac{Tr\left[\Sigma H(\mu)\right]}{2} \tag{16}$$

Proof Recall the objective of MFVI

$$L_{\text{MFVI}}(\mu) = L(\mu) + \big[\mathbb{E}_{\eta\sim\mathcal{N}(0,I)}L(\mu + \Sigma^{\frac{1}{2}}\eta) - L(\mu)\big], \tag{17}$$

where, similarly as in (Eqn. (1)), the term in brackets can be interpreted as a sharpness penalty. Note that we have fixed $\Sigma = \sigma^2 I$ for a constant σ. Using a second-order Taylor expansion around μ, where $H(\mu)$ is the Hessian at μ,

$$\begin{aligned}
L_{\text{MFVI}}(\mu) &\approx \mathbb{E}_{\eta\sim\mathcal{N}(0,I)}\big[L(\mu) + \eta^\top(\Sigma^{\frac{1}{2}})^\top\nabla_\theta L(\mu) + \frac{1}{2}\eta^\top(\Sigma^{\frac{1}{2}})^\top H(\mu)\Sigma^{\frac{1}{2}}\eta\big] && (18)\\
&= L(\mu) + \frac{1}{2}\mathbb{E}\Big[\text{Tr}(\eta^\top(\Sigma^{\frac{1}{2}})^\top H(\mu)\Sigma^{\frac{1}{2}}\eta)\Big] && (19)\\
&= L(\mu) + \frac{1}{2}\mathbb{E}\Big[\text{Tr}(\eta\eta^\top(\Sigma^{\frac{1}{2}})^\top H(\mu)\Sigma^{\frac{1}{2}})\Big] && (20)\\
&= L(\mu) + \frac{1}{2}\text{Tr}\Big[\mathbb{E}(\eta\eta^\top)(\Sigma^{\frac{1}{2}})^\top H(\mu)\Sigma^{\frac{1}{2}}\Big] && (21)\\
&= L(\mu) + \frac{1}{2}\text{Tr}\big[\Sigma H(\mu)\big]. && (22)
\end{aligned}$$

If $\Sigma = \frac{\rho^2}{p}I$, we recover the first part of the theorem. ■

Since the MFVI objective in (Eqn. (17)) is not typically available in closed form, in practice one uses a single-sample Monte Carlo estimate based on the reparametrization trick. This yields a stochastic objective as follows.

$$L_{\mathrm{MFVI}}(\mu) \approx L(\mu + \Sigma^{\frac{1}{2}}\eta), \quad \eta \sim \mathcal{N}(0, I) \tag{23}$$

We can follow the method of [?] to see that SGD on the single-sample Monte Carlo estimates of the gradient display additional implicit regularisation towards wider minima. This is carried out in Proposition 4.

Proposition 4 *Gradient descent with MFVI step follows a path that is closest to gradient flow path on $\tilde{L}_{MFVI}(\mu, \Sigma)$, where $\tilde{L}_{MFVI}(\mu, \Sigma)$ is the following*

$$\tilde{L}_{MFVI}(\mu, \Sigma) \approx \mathbb{E}_{\eta \sim \mathcal{N}(0,I)}\left[L(\mu + \Sigma^{\frac{1}{2}}\eta) + \frac{\delta}{4}||\nabla_\theta L(\mu + \Sigma^{\frac{1}{2}}\eta)||_2^2\right] \tag{24}$$

$$\approx L(\mu) + \frac{\delta}{4}||g(\mu)||_2^2 + \frac{\delta}{4}Tr[\Sigma H(\mu)^2] \tag{25}$$

where $g(\mu + \Sigma^{\frac{1}{2}}\eta) = \nabla_\theta L(\mu + \Sigma^{\frac{1}{2}}\eta)$.

Proof

By [1, 21], we have the first approximation, i.e.

$$\tilde{L}_{\mathrm{MFVI}}(\mu, \Sigma) \approx \mathbb{E}_{\eta \sim \mathcal{N}(0,I)}\left[L(\mu + \Sigma^{\frac{1}{2}}\eta) + \frac{\delta}{4}||\nabla_\theta L(\mu + \Sigma^{\frac{1}{2}}\eta)||_2^2\right] \tag{26}$$

Now with Taylor expansion we get

$$L(\mu + \Sigma^{\frac{1}{2}}\eta) = L(\mu) + (\Sigma^{\frac{1}{2}}\eta)^{\mathrm{T}} g(\mu) + O(||\Sigma^{\frac{1}{2}}\eta||_2^2) \tag{27}$$

and

$$g(\mu + \Sigma^{\frac{1}{2}}\eta) = g(\mu) + H(\mu)\Sigma^{\frac{1}{2}}\eta + O(||\Sigma^{\frac{1}{2}}\eta||_2^2) \tag{28}$$

Now by using (26), (27) and (28) we obtain

$$\tilde{L}_{\mathrm{MFVI}}(\mu, \Sigma) \approx \mathbb{E}_{\eta \sim \mathcal{N}(0,I)}\left[L(\mu) + (\Sigma^{\frac{1}{2}}\eta)^\top g(\mu) + \frac{\delta}{4}(g(\mu) + H(\mu)\Sigma^{\frac{1}{2}}\eta)^\top(g(\mu) + H(\mu)\Sigma^{\frac{1}{2}}\eta)\right] \tag{29}$$

$$= L(\mu) + \frac{\delta}{4}\mathbb{E}_{\eta \sim \mathcal{N}(0,I)}\left[g(\mu)^\top g(\mu) + 2g(\mu)^\top H(\mu)\Sigma^{\frac{1}{2}}\eta + (H(\mu)\Sigma^{\frac{1}{2}}\eta)^\top(H(\mu)\Sigma^{\frac{1}{2}}\eta)\right] \tag{30}$$

$$= L(\mu) + \frac{\delta}{4}g(\mu)^\top g(\mu) + \frac{\delta}{4}\mathbb{E}\left[\mathrm{Tr}\left[(H(\mu)\Sigma^{\frac{1}{2}}\eta)^\top(H(\mu)\Sigma^{\frac{1}{2}}\eta)\right]\right] \tag{31}$$

$$= L(\mu) + \frac{\delta}{4}g(\mu)^{\mathrm{T}} g(\mu) + \frac{\delta}{4}\mathbb{E}\left[\mathrm{Tr}\left[\eta\eta^{\mathrm{T}}\Sigma^{\frac{1}{2}}H(\mu)^2\Sigma^{\frac{1}{2}}\right]\right] \tag{32}$$

$$= L(\mu) + \frac{\delta}{4}||g(\mu)||_2^2 + \frac{\delta}{4}\mathrm{Tr}\left[\mathbb{E}\left[\eta\eta^{\mathrm{T}}\Sigma^{\frac{1}{2}}H(\mu)^2\Sigma^{\frac{1}{2}}\right]\right] \tag{33}$$

$$= L(\mu) + \frac{\delta}{4}||g(\mu)||_2^2 + \frac{\delta}{4}\mathrm{Tr}\left[\mathbb{E}\left[\eta\eta^{\mathrm{T}}\right]\Sigma^{\frac{1}{2}}H(\mu)^2\Sigma^{\frac{1}{2}}\right] \tag{34}$$

$$= L(\mu) + \frac{\delta}{4}||g(\mu)||_2^2 + \frac{\delta}{4}\mathrm{Tr}\left[\Sigma H(\mu)^2\right]. \tag{35}$$

■

This shows that Mean Field Variational Inference implicitly regularizes the trace of the square of the Hessian, and hence it also regularizes the magnitude of its eigenvalues. A way of defining sharpness is via the local curvature of the loss function around the minimum given that it is a critical point [4]. Since local curvature is encoded in the Hessian eigenvalues, this means that MFVI penalizes a notion of sharpness at critical points of the loss landscape

C. Algorithms

Algorithm 1 Random SAM Algorithm - Variational Optimization

```
Input     : Training set $S = \{(x_i, y_i)\}$, parameter $\sigma_0$
Initialize: $\mu$
for $t = 1, 2, \cdots$ do
    (1) Sample batch $B \sim S$
    (2) Take a sample of $\eta \sim N(0, 1)$
    (3) Compute the gradient of the loss of $\mu$ on batch B, i.e. $\left[\nabla_\mu L_B\right]_{\mu+\sigma_0\eta}$
    (4) Update $\mu$, i.e.
    $\mu \longleftarrow \mu - \eta_1 \left[\nabla_\mu L_B\right]_{\mu+\sigma_o\eta}$
end
```

Algorithm 2 Mean Field Variational Inference Algorithm

```
Input     : Training set $S = \{(x_i, y_i)\}$, parameter $\sigma_0$
Initialize: $\Sigma$ and $\mu$
for $t = 1, 2, \cdots$ do
    (1) Sample batch $B \sim S$
    (2) Take a sample of $\eta \sim N(0, 1)$
    (3) Compute the gradient of the loss of $\mu$ on batch B, i.e. $\left[\nabla_\mu L_B\right]_{\mu+\Sigma^{\frac{1}{2}}\eta}$
    (4) Compute the loss of $\Sigma$ on batch B, i.e. $\left[\nabla_\Sigma L_B\right]_{\mu+\Sigma^{\frac{1}{2}}\eta} + \nabla_\Sigma \mathrm{KL}[(\mu, \Sigma)||(0, \sigma_0 I)]$
    (5) Update $\mu$ and $\Sigma$, i.e.
    $\mu \longleftarrow \mu - \eta_1 \left[\nabla_\mu L_B\right]_{\mu+\Sigma^{\frac{1}{2}}\eta}$
    $\Sigma \longleftarrow \Sigma - \eta_2 \left(\nabla_\Sigma \left[L_B\right]_{\mu+\Sigma^{\frac{1}{2}}\eta} + \nabla_\Sigma \mathrm{KL}[(\mu, \Sigma)||(0, \sigma_0 I)]\right)$
end
```

Algorithm 3 VariationalSAM Algorithm

```
Input      : Training set $S = \{(x_i, y_i)\}$, parameters $\alpha$ and $\beta$,
             learning rates $\eta_1$ and $\eta_2$.
Initialize: $\Sigma$ and $\mu$
for $t = 1, 2, \cdots$ do
    (1) Sample batch $B \sim S$
    (2) Compute the gradient of the loss on batch B, i.e. $\nabla_\mu L_B(\mu)$
    (3) Compute $\epsilon^*_{\text{VSAM}}(\mu)$ using (2)
    (4) Compute gradient approximation for the VSAM loss,
    i.e. $\nabla_\mu L_{\text{VSAM}}(\mu, \Sigma) \approx \frac{\partial L}{\partial \mu}|_{\mu+\epsilon^*_{\text{VSAM}}(\mu)}$
    (5) Update $\mu \longleftarrow \mu - \eta_1 \frac{\partial L}{\partial \mu}|_{\mu+\epsilon^*_{\text{VSAM}}}$
    (6) Compute the gradient of VariationalSAM loss on batch B, i.e. $\nabla_\Sigma L_{\text{VSAM}}(\mu, \Sigma)$
    (7) Update $\Sigma \longleftarrow \Sigma - \eta_2 \nabla_\Sigma L_{\text{VSAM}}(\mu, \Sigma)$
end
```

D. Details of Experiments

We use WideResNets [24] on the CIFAR-10/100 datasets [14]. Following prior work [6, 12, 15] we calibrate the SGD optimiser with momentum 0.9, weight decay 0.0005 and initial learning rate 0.1. We use stepwise decreasing learning rate scheduling as we have found this more effective than cosine learning rate scheduling. Using batch size 128, we train the optimizers requiring two backpropagations per step (SAM, VSAM, MixSAM) for up to 200 epochs, while those with only one backpropagation (SGD and RandomSAM) are trained for up to 400 epochs. For CIFAR-10, we employ label smoothing [18] with factor 0.1.

We follow [6] in setting $\rho = 0.05$ in the SAM optimizer. Random SAM requires the standard deviation of the noise σ be specified. After exploration of the hyperparameter space, we have set this to $5e - 5$. In VariationalSAM, we used a learning rate of 0.01 on Σ. We set values for ρ and the penalty coefficients in order to match the KL-divergence of the prior and posterior distributions. The matrix Σ was taken to be diagonal and was initialized in order to render the perturbation size around 0.05. For details see Appendix E. The results are summarized in Table 3.

E. Information theoretic motivation for VariationalSAM

The SAM, ASAM and FSAM algorithms all modify the loss function to penalise the maximum loss value within a small neighbourhood around the current weights. The considered neighbourhood is an Euclidean ball in SAM, a weight dependent ellipsoid in ASAM, and an ellipsoid defined by the Fisher information matrix in FSAM. What we suggest is a generalization of these approaches: our method omits any constraint on the ellipsoid defining the neighbourhood, and treats it as an object to learn. Specifically, besides μ the algorithm also optimizes a symmetric positive-definite matrix Σ, because for such matrix $\epsilon^\top \Sigma \epsilon \leq p$ describes an ellipsoid.

Following the ideas of the previously mentioned algorithms, the modified loss function would be

$$L_{\text{VSAM}}(\mu, \Sigma) = \max_{\epsilon^\top \Sigma^{-1} \epsilon \leq p} L(\mu + \epsilon). \tag{36}$$

Our intention is to do coordinate descent on $L_{\text{VSAM}}(\mu, \Sigma)$ w.r.t. μ and Σ. However, minimizing the loss in Σ would lead to the null matrix, which is undesirable, therefore we would like to add additional terms to the loss. The idea is motivated by the Variational Bayesian methods, specifically the maximization of the Evidence Lower Bound (ELBO) and equivalently, the minimization of the negative ELBO. Consider a prior $p(Z)$, a likelihood $p(X|Z)$, and an arbitrary distribution $q_\theta(Z)$, then the posterior $p(Z|X)$ and the evidence $p(X)$ can be computed, and the minimization of the negative ELBO has the form of

$$\begin{aligned}\underset{\theta}{\operatorname{argmin}} -\mathcal{L}_{ELBO} &= \underset{\theta}{\operatorname{argmin}}\, \mathrm{KL}(q_\theta(Z)||p(Z|X)) - \log p(X) \\ &= \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{q_\theta} - \log p(X|Z) + \mathrm{KL}(q_\theta(Z)||p(Z))\end{aligned} \tag{37}$$

When the loss function is defined as the negative log-likelihood, the expectation can be considered as the expectation of the loss.

$$\underset{\theta}{\operatorname{argmin}} -\mathcal{L}_{ELBO} = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{z\sim q_\theta} L(z) + \frac{1}{N}\mathrm{KL}(q_\theta(Z)||p(Z)) \tag{38}$$

Choosing $p(Z) = \mathcal{N}(0, \sigma_0^2 I)$ and $q_\theta(Z) = \mathcal{N}(\mu, \Sigma)$ as k-dimensional Gaussians, the KL-divergence can be rewritten as

$$\mathrm{KL}\big[\mathcal{N}(\mu,\Sigma)||\mathcal{N}(0,\sigma_0^2 I)\big] = \frac{1}{2}\left[\frac{1}{\sigma_0^2}\mathrm{Tr}\Sigma + \log\det\Sigma^{-1} + \log\sigma_0^{2k} + \frac{1}{\sigma_0^2}\|\mu\|^2 - k\right]. \tag{39}$$

By slightly rephrasing $\mathbb{E}_{z\sim q_\mu} L(z)$ as $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma)} L(\mu+\epsilon)$ we get

$$\underset{\mu}{\operatorname{argmin}} -\mathcal{L}_{ELBO} = \underset{\mu}{\operatorname{argmin}}\, \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma)} L(\mu+\epsilon) + \frac{1}{2N\sigma_0^2}\mathrm{Tr}\Sigma + \frac{1}{2N}\log\det\Sigma^2 + \frac{1}{2N\sigma_0^2}\|\mu\|^2 \tag{40}$$

The expectation above can be bounded with $\max_{\epsilon^\top\Sigma^{-1}\epsilon\leq p} L(\mu+\epsilon)$ in a similar way to the derivations in SAM and FSAM, if ρ is sufficiently large. This motivates the minimization of

$$\max_{\epsilon^\top\Sigma^{-1}\epsilon\leq p} L(\mu+\epsilon) + \frac{1}{2N\sigma_0^2}\mathrm{Tr}\Sigma + \frac{1}{2N}\log\det\Sigma^{-1} + \frac{1}{2N\sigma_0^2}\|\mu\|^2. \tag{41}$$

We note that the derivations of SAM, ASAM, FSAM also use the same approach, but they bound the KL term further. Instead, we found these terms interesting to keep.

Using Appendix A, we can approximate the loss as

$$L_{\mathrm{VSAM}}(\mu,\Sigma) = L(\mu) + \sqrt{p}\sqrt{\nabla^\top l(\mu)\Sigma\nabla L(\mu)} + \frac{1}{2N\sigma_0^2}\mathrm{Tr}\Sigma + \frac{1}{2N}\log\det\Sigma^{-1} + \frac{1}{2N\sigma_0^2}\|\mu\|^2. \tag{42}$$

We take gradient descent steps w.r.t. μ and Σ alternately.

F. PAC-Bayes bound for VariationalSAM

Theorem 5 *For a parameter space $\mathcal{M}$ (with later described properties) and for any $(\mu, \Sigma) \in \mathcal{M} \times \Theta$ we have with probability at least $1-\delta$, that*

$$\mathbb{E}_{\epsilon \sim N(0,\Sigma)}[L_D(\mu+\epsilon)] \leq max_{\epsilon^T\Sigma^{-1}\epsilon \leq \gamma^2} L_S(\mu+\epsilon) + \sqrt{\frac{O(m+\log\frac{n}{\delta})}{n-1}} \tag{43}$$

where L_D is the generalization loss, and $max_{\epsilon^T\Sigma^{-1}\epsilon \leq \gamma^2} L_S(\mu+\epsilon)$ is the empirical SAM loss, with the geometry provided by Σ, i.e. the VSAM loss without the KL divergence term and $\gamma = m(1+\sqrt{\log n/m})$

Proof To build our PAC-Bayes bound we follow the steps laid out in the PAC-Bayes bound for Fisher SAM [12], and we omit some details which can be found there. Let's take the parameter spaces $\mu \in \mathcal{M} \subset \mathbb{R}^m$ and $\Sigma \in \Theta$, where Θ is the subset of the diagonal, positive definite matrices. Also assume, that $\text{diam}(\mathcal{M}) \leq M$ and $1/\lambda \leq |\Sigma_{k,k}| \leq \lambda$ for any $\Sigma \in \Theta$ and k. We will take the set of ellipsoids $P_{1,1}, \cdots, P_{t,s}$, where

$$P_{i,j} = \{(\mu,\Sigma) \in \mathcal{M} \times \Theta : (\mu-\mu_i)^T \Sigma_j^{-1}(\mu-\mu_i) \leq r^2 \text{ and } \frac{[\Sigma_j]_{k,k}}{\Sigma_{k,k}} = 1 + c_k, \tag{44}$$

$$\text{where } c_k \in [-c, c] \quad \forall k\} \tag{45}$$

There exists a finite set of these ellipsoids, that cover $\mathcal{M} \times \Theta$.

1. Let's take the following subset of $\mathbb{R}^m$

$$P_{i,j}|_\mu = \{\mu : \mu \in P_{i,j}\}$$
$$vol(P_{i,j}|_\mu) \propto r^m \times \Sigma_j^{\frac{1}{2}}$$

   thus $t = O(M^m/r^m) =$, so $\log(t) = O(m)$.

2. We also assumed $\frac{1}{\lambda} \leq \Sigma_{k,k} \leq \lambda$. Let's think about Θ as a subset of $\mathbb{R}^m$. Then as in part (1) we can define

$$P_{i,j}|_\Sigma = \{\Sigma : \Sigma \in P_{i,j}\}$$
$$vol(P_{i,j}|_\Sigma) \propto (2c)^m$$

   thus $s = O((\lambda + \frac{1}{\lambda})^m/(2c)^m)$, so $\log(s) = O(m)$

Now we use the PAC-Bayes Theorem of McAllester [17] as follows. For any prior distribution $P(\mu, \Sigma)$, posterior distribution $Q(\mu, \Sigma)$ and training set S we have with probability at least $1-\delta$

$$\mathbb{E}_{Q(\mu,\Sigma)}[L_D(\mu,\Sigma)] \leq \mathbb{E}_{Q(\mu,\Sigma)}[L_S(\mu,\Sigma)] + \sqrt{\frac{\text{KL}[Q(\mu,\Sigma)||P(\mu,\Sigma)] + \log\frac{n}{\delta}}{2(n-1)}} \tag{46}$$

$L_D(\mu, \Sigma)$ and $L_S(\mu, \Sigma)$ are generalization and empirical losses. We choose $Q(\mu, \Sigma) = N(\mu_0, \Sigma_0)$, where $(\mu_0, \Sigma_0) \in \mathcal{M} \times \Theta$ and choose our set of priors as $P_{1,1}, \cdots, P_{t,s}$, where $P_{i,j} = N(\mu_i, \Sigma_j)$. Then as in (46) for each i, j we have with probability $1 - \delta_{i,j}$, that

$$\forall Q(\mu, \Sigma) \quad \mathbb{E}_{Q(\mu,\Sigma)}[L_D(\mu, \Sigma)] \leq \mathbb{E}_{Q(\mu,\Sigma)}[L_S(\mu, \Sigma)] + \sqrt{\frac{\text{KL}[Q(\mu, \Sigma)||P_{i,j}(\mu, \Sigma)] + \log \frac{n}{\delta_{i,j}}}{2(n-1)}} \tag{47}$$

In the intersection (47) holds for every $P_{i,j}$. So by Union bound theorem the intersection is at least $1 - \sum_{i,j} \delta_{i,j}$, so if we take $\delta_{i,j} = \frac{\delta}{st}$, we have with probability at least $1 - \delta$

$$\forall Q(\mu, \Sigma) \quad \mathbb{E}_{Q(\mu,\Sigma)}[L_D(\mu, \Sigma)] \leq \mathbb{E}_{Q(\mu,\Sigma)}[L_S(\mu, \Sigma)] + \tag{48}$$

$$+\sqrt{\frac{\text{KL}[Q(\mu, \Sigma)||P_{i,j}(\mu, \Sigma)] + \log \frac{n}{\delta} + \log(s) + \log(t)}{2(n-1)}} \quad \forall i, j \tag{49}$$

If we choose the prior closes to $Q(\mu, \Sigma)$ we get the following:

$$\text{KL}[Q||P_{i,j}] = \frac{1}{2}\left(\text{Tr}(\Sigma_j \Sigma_0^{-1}) + (\mu_0 - \mu_i)^T \Sigma_j^{-1}(\mu_0 - \mu_i) + \log \frac{|\Sigma_0|}{|\Sigma_j|} - m\right) \tag{50}$$

From our assumptions $\text{Tr}(\Sigma_j \Sigma_0^{-1}) \leq m(1+c)$, $\log \frac{|\Sigma_0|}{|\Sigma_j|} \leq \sum_k log(1+c_k) \leq \sum_k c_k \leq mc$ and $(\mu_0 - \mu_i)^T \Sigma_j^{-1}(\mu_0 - \mu_i) \leq r^2$. Thus we get

$$\text{KL}[Q||P_{i,j}] \leq \frac{1}{2}(m + mc + r^2 + mc - m) = mc + \frac{r^2}{2} \tag{51}$$

Lets rephrase 48 as we plug in $\mu + \epsilon$ instead of μ where ϵ $N(0, \Sigma)$. Thus we get:

$$\forall \mu, \Sigma \quad \mathbb{E}_{\epsilon \sim N(0,\Sigma)}[L_D(\mu + \epsilon)] \leq \mathbb{E}_{\epsilon \sim N(0,\Sigma)}[L_S(\mu + \epsilon)] + \sqrt{\frac{mc + r^2/2 + \log \frac{n}{\delta} + \log(s) + \log(t)}{2(n-1)}} \tag{52}$$

Let $u = \Sigma^{-1/2}\epsilon$, so $u \sim N(0, 1)$ using the result from Laurent & Massart [16] we get that with probability at least $1 - \frac{1}{\sqrt{n}}$

$$|u|_2^2 = \epsilon^T \Sigma^{-1} \epsilon \leq m(1 + \sqrt{\log n/m})^2 = \gamma^2 \tag{53}$$

Let's partition the space into two parts. One where (53) holds, where we take the maximum loss and on where (53) does not hold, where we choose the loss bound L_{max}. So we have

$$\mathbb{E}_{\epsilon\sim N(0,\Sigma)}[L_S(\mu+\epsilon)] \leq (1-1/\sqrt{n})\text{max}_{\epsilon^T\Sigma^{-1}\epsilon\leq\gamma^2} L_S(\mu+\epsilon) + \frac{L_{\text{max}}}{\sqrt{n}} = \tag{54}$$

$$= \text{max}_{\epsilon^T\Sigma^{-1}\epsilon\leq\gamma^2} L_S(\mu+\epsilon) + \frac{L_{\text{max}}}{\sqrt{n}} \tag{55}$$

Plugging (54) into (52) we get:

$$\mathbb{E}_{\epsilon\sim N(0,\Sigma)}[L_D(\mu+\epsilon)] \leq \text{max}_{\epsilon^T\Sigma^{-1}\epsilon\leq\gamma^2} L_S(\mu+\epsilon) + \frac{L_{\text{max}}}{\sqrt{n}} + \tag{56}$$

$$+\sqrt{\frac{\frac{r^2(\sqrt{m}+\sqrt{\log n})^2}{2\gamma^2} + mc + \log\frac{n}{\delta} + \log(st)}{2(n-1)}} = \text{max}_{\epsilon^T\Sigma^{-1}\epsilon\leq\gamma^2} L_S(\mu+\epsilon) + \sqrt{\frac{O(m+\log\frac{n}{\delta})}{n-1}} \tag{57}$$

■

Note, that similar bounds can be build for Random SAM and MFVI.

G. Further Experiments

	WideResNet 28-2
SGD	$95.90^{\pm 0.07}$
SAM	$96.10^{\pm 0.11}$
ASAM	$96.17^{\pm 0.07}$
FSAM	$96.23^{\pm 0.06}$
VSAM	$94.18^{\pm 0.11}$
RSAM	$96.08^{\pm 0.13}$

Table 4: CIFAR-10

For these experiments, we have set additional hyperparameters to the value reported in [15] and [12]. Namely, in ASAM we use $\gamma = 0.5, \eta = 0.01$ for CIFAR-10, $\gamma = 1.0, \eta = 0.1$ for CIFAR-100 and in FSAM $\gamma = 0.1, \eta = 1.0$ for both datasets.