Formalized Axioms and Analysis

Formalizability Index: 0.36

Total Segments: 464

Formalizable Segments: 165

Logic Reconstruction:

=== Formal Logic Reconstruction ===
(0) Sharpness-aware minimization (SAM) is a promising method to improve generalization.
Formal: SAM is a promising method to enhance generalization.
(1) The original proposal of SAM by Foret et al. signifies a foundational shift.
Formal: Foret et al.'s proposal of SAM marks a fundamental change.
(4) We study SAM for out-of-distribution (OOD) generalization.
Formal: SAM is studied for OOD generalization.
(5) The original SAM outperforms the Adam baseline by 4.76% in zero-shot out-of-distribution (OOD) g
Formal: Original SAM surpasses Adam by 4.76% in zero-shot OOD generalization.
(6) An OOD (out-of-distribution) generalization bound can be provided in terms of sharpness for a part
Formal: OOD generalization can be bounded by sharpness in a specific context.
(7) Gradual domain adaptation (GDA) is a form of out-of-distribution (OOD) generalization that utilizes
Formal: GDA is OOD generalization using intermediate domains between source and target domain
(8) The original SAM outperforms the baseline of Adam on each of the experimental datasets by 0.82%
Formal: Original SAM exceeds Adam by 0.82% on average across datasets.
(8) The strongest SAM variants outperform Adam by 1.52% on average.
Formal: Top SAM variants outperform Adam by 1.52% on average.
(9) A generalization bound for SAM is provided in the GDA setting.
Formal: SAM has a generalization bound in GDA settings.
(10) Asymptotically, this generalization bound is no better than the one for self-training in the literature
Formal: This asymptotic generalization bound matches self-training in GDA literature.
(11) There is a disconnection between the theoretical justification for SAM and its empirical performan
Formal: Theoretical SAM justification and empirical performance are disconnected.
(12) Low sharpness alone does not account for all of SAM’s generalization benefits.
Formal: SAM's generalization benefits aren't solely due to low sharpness.
(13) There are potential avenues for obtaining a tighter analysis for Stochastic Approximation Methods
Formal: Potential exists for tighter SAM analysis in OOD context.
(14) Theoretical results provide a solid starting point for analyzing SAM in OOD settings.
Formal: Theoretical results are a basis for analyzing SAM in OOD.
(14) SAM can be applied to OOD settings to significantly improve accuracy.
Formal: SAM can enhance accuracy in OOD settings.
(14) Newer variants of SAM can be leveraged for further improvements in accuracy.
Formal: New SAM variants can further boost accuracy.
(15) Sharpness-Aware Minimization (SAM) is a promising new optimization algorithm.
Formal: SAM is an innovative optimization algorithm.
(16) A robust optimization procedure can lead to significant performance gains in the i.i.d.

Formal: Robust optimization can yield significant i.i.d. performance gains.

(18) SAM remains understudied in the out-of-distribution (OOD) generalization setting.
Formal: SAM is still understudied in OOD generalization.

(19) A number of SAM variants have been proposed to improve the accuracy and efficiency of the orig
Formal: Several SAM variants aim to boost accuracy and efficiency.

(22) The potential to enhance OOD (out-of-distribution) generalization exists.
Formal: Enhancing OOD generalization potential is present.

(24) There are eight SAM variants, including the original SAM.
Formal: Eight SAM variants exist, original SAM included.

(27) SAM can be used to improve zero-shot OOD generalization.
Formal: SAM improves zero-shot OOD generalization.

(27) The strongest SAM variants can be used for an even further improvement.
Formal: Top SAM variants achieve further improvements.

(29) A theoretical analysis of SAM under the distribution shift setting provides performance gains.
Formal: Theoretical SAM analysis during distribution shift yields gains.

(31) We extend the setting to gradual domain adaptation.
Formal: The setting is extended to gradual domain adaptation.

(32) SAM outperforms the Adam baseline by 0.82% on average.
Formal: SAM surpasses Adam by 0.82% on average.

(32) The strongest SAM variants achieve an even greater 1.52% average improvement over Adam.
Formal: Top SAM variants exceed Adam by 1.52% on average.

(32) SAM and the strongest SAM variants can be used for consistent performance gains in GDA.
Formal: Use SAM and top variants for consistent GDA gains.

(36) Our work is asymptotically the same as prior work in the GDA literature (Wang et al., 2022).
Formal: Our work matches prior GDA literature asymptotically.

(42) We further define the sharpness of θ and the corresponding ρ-robust empirical loss.
Formal: Sharpness of θ and ρ-robust empirical loss are defined.

(44) The ρ-robust risk of parameters θ ∈ Θ is the maximum loss obtained by perturbing θ in the worst p
Formal: ρ-robust risk is max loss from perturbing θ with ■2-norm ≤ ρ.

(45) The ρ-sharpness of parameters θ measures how much the loss increases when we perturb them
Formal: ρ-sharpness is loss increase from worst perturbation of θ with ■2-norm ≤ ρ.

(46) Sharpness-aware minimization (SAM) proposes minimizing the ρ-robust empirical loss rather thar
Formal: SAM minimizes ρ-robust empirical loss instead of standard loss.

(47) SAM’s objective is to find a minimizer θ■ of the form: θ■=argminθ∈Θ max β:■β■2≤ρ E(θ+β).
Formal: SAM aims to find θ■ = argminθ∈Θ max β:■β■2≤ρ E(θ+β).

(49) The SAM gradient drops a second-order term arising from the chain rule.
Formal: SAM gradient omits a chain rule second-order term.

(51) Our results can be extended to any ■p-norm.
Formal: Results extend to any ■p-norm.

(53) This should not be confused with m-sharpness from Foret et al.
Formal: Do not confuse with Foret et al.'s m-sharpness.

(55) SAM (Sharpness-Aware Minimization) is a method for optimizing model parameters.
Formal: SAM optimizes model parameters.

(55) SAM uses variant-specific oracles for gradient computation, perturbation, and descent steps.
Formal: SAM employs specific oracles for gradient, perturbation, and descent.

(57) Adaptive SAM (ASAM) is a version of SAM that uses a scale-invariant version of the first-order ap

Formal: ASAM is a SAM variant using scale-invariant first-order approximation.

(59) Adaptive SAM has one hyperparameter ρ.

Formal: ASAM includes one hyperparameter ρ.

(60) FisherSAM is a special case of ASAM with a specific normalization operator.

Formal: FisherSAM is an ASAM variant with a distinct normalization operator.

(62) K-SAM is a variant of SAM that only uses the top K data samples with the highest loss for gradier

Formal: K-SAM uses top K loss samples for gradient calculations.

(65) The idea behind Equation (12) is to remove the component of the descent gradient $\nabla \theta E(\theta)|\theta+\beta$■ly

Formal: Equation (12) eliminates descent gradient component paralleling the ascent gradient.

(67) Using only the full-batch direction of the ascent-step gradient impairs performance.

Formal: Reliance on full-batch ascent-step gradient direction hurts performance.

(69) Computing the full-batch gradient $\nabla \theta E(\theta)$ is computationally prohibitive.

Formal: Full-batch gradient calculation is computationally demanding.

(70) FriendlySAM uses a specific perturbation method defined by an equation.

Formal: FriendlySAM employs equation-defined perturbation.

(72) NoSAM is a variant of SAM that only performs the SAM perturbation on normalization layers.

Formal: NoSAM applies SAM perturbation solely on normalization layers.

(74) The computational cost is compared across three specific metrics.

Formal: Cost is compared using three distinct metrics.

(77) EfficientSAM (ESAM) is a variant of SAM intended to make SAM more efficient.

Formal: ESAM is a SAM variant for enhanced efficiency.

(78) Stochastic weight perturbation is applied by performing the SAM perturbation on a fraction of the v

Formal: SAM perturbation involves stochastic weight alteration.

(79) Sharpness-sensitive data selection is performed by only computing gradients over data samples v

Formal: Select sharpness-sensitive data by targeting highest loss increasing samples.

(83) A model θS is trained on a source domain $S \in \Delta(X \times Y)$ with a training set.

Formal: Train model θS on source domain S with training set.

(86) The zero-shot OOD generalization error is given by $ET(\theta S)$.

Formal: Zero-shot OOD error is $ET(\theta S)$.

(92) The intermediate domains are evenly distributed between source and target.

Formal: Intermediate domains are equally spaced between source and target.

(95) The intermediate domains are also evenly distributed between source and target.

Formal: Intermediate domains are evenly placed between source and target.

(112) Variants such as LookSAM, F-SAM, and FisherSAM offer the strongest and most consistent imp

Formal: LookSAM, F-SAM, FisherSAM show the best and consistent SAM improvements.

(113) LookSAM and NoSAM perform the best among the variants with reduced computational cost.

Formal: LookSAM and NoSAM excel among low computational cost variants.

(113) LookSAM and NoSAM often outperform the original SAM in addition to being more efficient.

Formal: LookSAM, NoSAM surpass original SAM and are more efficient.

(117) FisherSAM takes the information geometry of the data into account, using an approximation of th

Formal: FisherSAM uses data's information geometry to find max perturbation with Fisher matrix.

(118) Under the cross-entropy loss used specifically in the experiments in Table 2, the Fisher informat

Formal: In Table 2 experiments, Fisher matrix equals cross-entropy Hessian.

(119) FisherSAM could be understood as a second-order approximation of the loss perturbation.

Formal: FisherSAM is a second-order loss perturbation approximation.

(120) Unlike FisherSAM, the connection between the FriendlySAM objective and OOD performance is

Formal: FriendlySAM's OOD link is less explicit than FisherSAM's.

(123) The modified perturbation makes FriendlySAM significantly more robust to the choice of the ρ hy
Formal: FriendlySAM's modified perturbation enhances robustness to ρ selection.

(124) The optimal value of ρ depends intricately on the choice of dataset.
Formal: Optimal ρ value is dataset-dependent.

(124) FriendlySAM provides performance gains in experiments by penalizing sharpness adaptively and
Formal: FriendlySAM aids performance by adaptively, stably penalizing sharpness.

(126) The Wasserstein distance is the smallest cost of moving mass between two distributions measu
Formal: Wasserstein distance is minimal cost of redistributing between two distributions.

(127) The loss functions considered in this paper are Lipschitz continuous.
Formal: Considered loss functions are Lipschitz continuous.

(127) For the loss function ■, there exist constants ρ1, ρ2, ρ3 such that certain inequalities involving ■
Formal: For ■, constants ρ1, ρ2, ρ3 satisfy certain inequalities.

(130) This result holds for any choice of μ,ν on Y×X.
Formal: Result applies to any μ,ν on Y×X.

(131) Lemma 1 (Sharpness-Aware Error Difference Over Shifted Domains) posits a bound on the diffe
Formal: Lemma 1 bounds error difference between two distributions.

(132) Population error of a model θ can be bounded in terms of its empirical sharpness.
Formal: Model θ's population error is bound by its empirical sharpness.

(133) For any model θ ∈ Θ satisfying E(θ) ≤ E■~N(0,ρ2I)[E(θ+■)] for some ρ > 0, with high probability
Formal: For θ∈Θ, E(θ) ≤ E■~N(0,ρ2I)[E(θ+■)] implies E(θ) ≤ ˆEρ(θ) + O((k ln(■θ■2/ρ2) + ln(n/δ))/n)

(133) Using the error difference lemma from Lemma 1 and the PAC-Bayesian bound from Lemma 2, a
Formal: Error difference lemma and PAC-Bayesian bound offer OOD generalization bound for SAM

(134) This result upper bounds the error of a model θμ on domain μ by the error of θν on domain ν, the
Formal: This result bounds θμ's error on μ using θν's error on ν, PAC complexity, θμ sharpness, θμ-

(135) Theorem 1 (Sharpness-Aware Domain Adaptation Error): For distributions μ and ν over X×Y and
Formal: Theorem 1: Error on θμ bound by θν error and extra terms for distributions μ,ν.

(141) Each domain t∈[T] is a distribution μt over X×Y.
Formal: Each domain t ∈ [T] is distribution μt on X×Y.

(145) The distribution shift between successive pairs of gradually shifted distributions and the average
Formal: Successive and average distribution shifts are mathematically defined.

(145) In gradual domain adaptation, a learner progressively accesses unlabeled examples from interm
Formal: Gradual DA uses intermediate domains for minimizing target domain's error.

(146) Adding Gaussian perturbation around θμ increases the expected loss.
Formal: Gaussian perturbation near θμ raises expected loss.

(155) We choose the optimal number of intermediate domains T■ for SAM from Figure 1.
Formal: Optimal intermediate domains T■ for SAM selected from Figure 1.

(158) SAM can be applied to GDA to consistently achieve stronger performance.
Formal: SAM consistently boosts GDA performance.

(159) SAM is not too sensitive to the perturbation radius hyperparameter.
Formal: SAM's sensitivity to perturbation radius is minimal.

(160) SAM with ρ= 0.2 leads to the strongest performance on Rotated MNIST, Portraits, and Color MN
Formal: Using ρ= 0.2 achieves top SAM performance on Rotated MNIST, Portraits, Color MNIST.

(160) SAM with ρ= 0.05 leads to the strongest performance on Covertype.
Formal: Using ρ= 0.05 achieves top SAM performance on Covertype.

(161) The strongest SAM variants lead to an average improvement of 1.42% compared to using Adam

Formal: Top SAM variants outperform Adam by 1.42% on average.

(163) The SAM variants with reduced computational cost tend to underperform Adam on GDA.
Formal: Low-cost SAM variants underperform Adam in GDA.

(164) FriendlySAM outperforms SAM by 0.40% on average across all datasets.
Formal: FriendlySAM exceeds SAM by 0.40% on average over datasets.

(164) FisherSAM slightly underperforms SAM by 0.01%.
Formal: FisherSAM is 0.01% less efficient than SAM.

(165) SAM can be used to consistently improve target domain error for GDA.
Formal: SAM consistently reduces target domain error in GDA.

(166) The segment claims that there is an extension of Theorem 1 to the setting of gradual domain ada
Formal: Theorem 1 is extended for GDA.

(166) GDA is presented as a technique which improves target domain error by performing gradual self
Formal: GDA uses gradual self-training on unlabeled intermediates to improve target error.

(173) The discrepancy measure captures the non-stationarity of the gradually shifting data.
Formal: Discrepancy measure reflects data's non-stationarity.

(174) The sequential Rademacher complexity generalizes the standard Rademacher complexity to the
Formal: Sequential Rademacher complexity adapts standard complexity for online learning.

(176) Each of the nT samples is viewed as the smallest element of the adaptation process.
Formal: nT samples are smallest adaptation process elements.

(178) Generalization bound can be stated for GDA performed using SAM.
Formal: Generalization bound exists for SAM-used GDA.

(179) The population risk of the gradually adapted model θT can be bounded under certain conditions.
Formal: Population risk for gradually adapted θT is boundable.

(182) By applying Corollary 2 of Kuznetsov & Mohri (2020a), a preliminary bound on the error in the ta
Formal: Error bound for target domain ET(θT) obtained using Kuznetsov & Mohri.

(184) The bound we obtain in Theorem 2 is of a specific mathematical form.
Formal: Theorem 2 offers mathematically specific bound.

(189) Adding Gaussian perturbation to each solution θt increases the expected loss.
Formal: Each θt with Gaussian perturbation raises expected loss.

(189) Theorem 1 must relate the parameters of the successive domains separately.
Formal: Theorem 1 requires separate parameter relationships.

(190) The PAC Bayes bound introduces an additional weight norm term W_avg.
Formal: PAC Bayes bound adds W_avg weight norm term.

(191) The PAC Bayesian bound yields a different sample complexity term compared to the original ana
Formal: PAC Bayesian bound differs from Rademacher in sample complexity.

(197) The overall sample complexity is expected to remain the same between our Theorem 2 and the
Formal: Our Theorem 2 and Wang et al. expect consistent sample complexity.

(200) A tighter error difference between shifted domains would improve the analysis.
Formal: Tighter shifted domain error difference enhances analysis.

(201) A localized analysis could exploit implicit properties of SAM.
Formal: Localized analysis can leverage SAM implicit properties.

(203) There is a relationship between sharpness and generalization.
Formal: Sharpness is related to generalization.

(206) Sharpness is among the empirical measures most strongly correlated with generalization.
Formal: Sharpness strongly correlates with generalization empirically.

(208) Sharp minima can generalize under reparameterizations that cause flat minima to become arbitr

Formal: Sharp minima can generalize despite reparameterizations causing flat minima sharpness.
(209) A measure of sharpness is tied to the information geometry of the data and is invariant under rep
Formal: Sharpness measure is data geometry-related and reparameterization-invariant.
(210) Many works have explored algorithms that lead to flatter solutions.
Formal: Research explores algorithms creating flatter solutions.
(211) In the context of domain generalization (DG), sharpness affects generalization.
Formal: In DG, sharpness impacts generalization.
(212) A modified version of stochastic weight averaging leads to flatter minima with improved DG.
Formal: Modified stochastic weight averaging results in flatter minima, improving DG.
(213) Generalization bounds depend on the empirical robust loss in the source domain.
Formal: Generalization bounds rely on source domain empirical robust loss.
(215) A flatness-aware minimization algorithm for DG leads to improved performance.
Formal: Flatness-aware minimization enhances DG performance.
(217) Out-of-distribution (OOD) generalization bounds are presented based on sharpness.
Formal: OOD generalization bounds are given in terms of sharpness.
(219) Sharpness-Aware Minimization (SAM) was originally proposed in Foret et al.
Formal: SAM was initially suggested by Foret et al.
(220) A PAC Bayesian analysis provides a generalization bound in terms of the expected sharpness o
Formal: PAC Bayesian analysis offers a bound based on expected sharpness with isotropic Gaussia
(222) The practical implementation of SAM uses this first-order approximation.
Formal: SAM's practical use employs first-order approximation.
(224) The flatness of the final solution does not sufficiently capture the generalization benefit from SAM
Formal: Final solution flatness inadequately captures SAM's generalization benefit.
(225) SAM leads to lower rank features with fewer active ReLU units.
Formal: SAM reduces feature rank and ReLU unit activity.
(225) SAM enhances feature quality by selecting more balanced features.
Formal: SAM boosts feature quality by balancing selection.
(225) SAM enhances robustness to label noise through implicitly regularizing the model Jacobian.
Formal: SAM enhances label noise robustness by implicit Jacobian regularization.
(225) SAM has an implicit denoising mechanism which prevents harmful overfitting in settings when S
Formal: SAM's implicit denoising deters harmful overfitting, unlike SGD.
(226) Many variants of SAM have been proposed to improve the efficiency and accuracy of the origina
Formal: SAM variants aim to improve original SAM efficiency and accuracy.
(229) Gradual self-training (GST) in GDA outperforms standard self-training without intermediate doma
Formal: GST in GDA overshadows standard self-training lacking intermediate domains.
(230) These bounds have an exponential dependence on the number of intermediate domains T.
Formal: Bounds exponentially depend on intermediate domains T.
(234) The analysis can be generalized to any ρ-Lipschitz losses and Wasserstein distances of any ord
Formal: Analysis generalizes to any ρ-Lipschitz losses and $p \geq 1$ Wasserstein distances.
(235) The existence of an optimal choice of intermediate domains T is suggested by the refined bound
Formal: Refined bounds imply optimal intermediate domains T exist.
(238) A new method of generating intermediate domains in an encoded feature space is proposed.
Formal: New method proposed for creating intermediate domains in encoded feature space.
(242) The main limitation of this work is the discrepancy between our theoretical analysis based on sha
Formal: Major limitation is discrepancy in sharpness-based theoretical analysis and prior work's asy
(244) The analysis for SAM can be tightened.

Formal: SAM analysis can be refined.
(247) SAM contributes to out-of-distribution generalization.
Formal: SAM aids OOD generalization.
(248) The original SAM achieved a 4.76% average improvement over the Adam baseline.
Formal: Original SAM improves Adam by 4.76% on average.
(248) The strongest SAM variants achieved an 8.01% average improvement over the Adam baseline.
Formal: Top SAM variants top Adam by 8.01% on average.
(249) An OOD (Out-of-Distribution) generalization bound can be derived based on sharpness.
Formal: OOD generalization bound is derivable from sharpness.
(252) We provided an extension of our OOD generalization bound to get a generalization bound based
Formal: We've extended our OOD bound to a sharpness-based generalization bound for GDA.
(254) There is a discrepancy between theoretical and empirical results regarding SAM.
Formal: Discrepancy exists between SAM's theoretical and empirical results.
(255) Our theoretical results provide a starting point for doing this.
Formal: Theoretical results offer a starting point.
(255) Our empirical results suggest that SAM can be used empirically to achieve significant gains for O
Formal: Empirical results indicate SAM's OOD generalization efficacy.
(257) Sharpness-aware minimization leads to low-rank features.
Formal: SAM results in low-rank features.
(265) There is a cohesive theory that addresses how learning occurs across different domains.
Formal: A cohesive theory explains multi-domain learning.
(276) Domain generalization can be achieved by seeking flat minima.
Formal: Seek flat minima for domain generalization.
(279) Entropy-sgd biases gradient descent into wide valleys.
Formal: Entropy-sgd directs gradient descent to wide valleys.
(282) Sharpness-aware minimization generalizes better than SGD.
Formal: SAM generalizes better than SGD.
(285) Sharp minima can generalize for deep nets.
Formal: Sharp minima generalize in deep networks.
(293) Efficient sharpness-aware minimization leads to improved training of neural networks.
Formal: Efficient SAM enhances neural network training.
(296) Sharpness-aware minimization efficiently improves generalization.
Formal: SAM efficiently boosts generalization.
(305) Flat minima in the context of optimization and machine learning refer to regions in the parameter
Formal: Flat minima in optimization offer invariant regions for better generalization.
(311) Batch normalization accelerates deep network training.
Formal: Batch normalization speeds up deep network training.
(314) Averaging weights leads to wider optima and better generalization.
Formal: Weight averaging provides wider optima for better generalization.
(321) Large-batch training for deep learning leads to a generalization gap.
Formal: Large-batch training induces a generalization gap.
(321) Large-batch training results in sharp minima.
Formal: Sharp minima arise from large-batch training.
(329) Information geometry provides a framework for understanding optimization techniques like sharp
Formal: Information geometry explains SAM optimization.
(335) Self-training is an effective method for gradual domain adaptation.

Formal: Self-training effectively facilitates gradual domain adaptation.
(340) Discrepancy-based theory is useful for forecasting non-stationary time series.
Formal: Discrepancy theory aids in non-stationary time series prediction.
(346) Adaptive sharpness-aware minimization (Asam) is proposed for scale-invariant learning in deep
Formal: ASAM supports scale-invariant deep neural network learning.
(368) The natural gradient method provides new insights and perspectives.
Formal: Natural gradient method offers fresh insights.
(372) Normalization layers are all that sharpness-aware minimization needs.
Formal: SAM requires only normalization layers.
(378) Online learning can be understood through the framework of sequential complexities.
Formal: Sequential complexities elucidate online learning.
(384) Sharpness-aware minimization enhances feature quality via balanced learning.
Formal: SAM improves feature quality through balanced learning.
(388) Dropout is a simple way to prevent neural networks from overfitting.
Formal: Dropout prevents neural network overfitting.
(392) Gradual domain adaptation can be better understood through improved analysis.
Formal: Enhanced analysis clarifies gradual domain adaptation.
(396) Sharpness minimization algorithms do not only minimize sharpness to achieve better generaliza
Formal: Sharpness algorithms use more than minimizing sharpness for generalization.
(402) Averaging weights of multiple fine-tuned models improves accuracy without increasing inference
Formal: Weight averaging of fine-tuned models enhances accuracy without extra inference time.
(407) A theoretical framework for out-of-distribution generalization is necessary.
Formal: OOD generalization requires a theoretical framework.
(410) Flatness-aware minimization is a crucial method for improving domain generalization.
Formal: Flatness-aware minimization crucially improves domain generalization.
(414) Invariant representations are crucial for effective domain adaptation.
Formal: Invariant representations are key to domain adaptation.
(419) There are fundamental limits in invariant representation learning.
Formal: Invariant representation learning has fundamental limits.
(419) Tradeoffs exist in the process of learning invariant representations.
Formal: Invariant representation learning involves tradeoffs.
(422) Gradual domain adaptation via gradient flow is a viable method.
Formal: GDA by gradient flow is viable.
(425) Robust out-of-distribution generalization can be achieved through considerations of sharpness.
Formal: Sharpness considerations secure robust OOD generalization.
(428) |Eρμ(θμ)–Eν(θν)|≤Sρ(θμ) +O(■θμ–θν■+Wp(μ,ν))
Formal: Error difference bound by sharpness and distribution distance.
(430) Given distributions μ, ν over X × Y and an error function E with loss satisfying Assumption 1 with
Formal: Eμ(θμ) is error-bounded by Eν(θν) with high probability, based on Assumption 1.
(430) A sharpness-aware generalization bound, along with Rademacher complexity and robust error d
Formal: Sharpness-aware bound with Rademacher and robust error provides error expectation bou
(431) If E(θ) ≤ E■~N(0,ρ2I)[E(θ+■)], then with probability ≥ 1–δ, E(θ) ≤ ˆEρ(θ) + O(sqrt((k ln(■θ■2/ρ2)
Formal: If E(θ) ≤ E■~N(0,ρ2I)[E(θ+■)], E(θ) is bounded with high probability.
(433) Theorem 2 is a key claim that is restated within the context of 'Total Sharpness-Aware Error Und
Formal: Theorem 2 restates 'Total Sharpness-Aware Error Under GDA'.
(434) The population risk of the gradually adapted model θT can be bounded with high probability.

Formal: Population risk of gradually adapted θT is bounded with high probability.

(438) The term ET(θT) can be bounded by a sequence of steps applying Theorem 1.

Formal: ET(θT) is bounded using Theorem 1 in steps.

(438) ET–1(θT) and other successive terms (ET–2(θT), ET–3(θT), etc.) can be bound similarly using th

Formal: Terms like ET–1(θT) are similarly bound using derived formulas.

(439) The text provides a bound on the expected value ET(θT) given initial conditions and average val

Formal: Bound on ET(θT) uses initial conditions, average parameter values over T.

(442) |Eµ(θ)–Eν(θ)|≤O (Wp(µ,ν))

Formal: Error difference limited by Wasserstein distance.

(444) Proposition 1 (Discrepancy Bound - Lemma 2 of Wang et al.

Formal: Proposition 1 gives discrepancy bound per Wang et al.'s Lemma 2.

(446) disc(qt) ≤ O/parenleft| iggt–1/summationdisplay k=0 qk(t–k–1)Wp(µk,µk+1)/parenright| igg

Formal: disc(qt) is bounded by distribution shifts.

(446) disc(qt) ≤ O(t∆) when qt=q■ t:= (1/t,..., 1/t)

Formal: disc(qt) follows O(t∆) under specific conditions.

(448) Definition 8 (Rademacher Complexity) introduces the concept of empirical Rademacher comple

Formal: Rademacher Complexity defines empirical complexity for models.

(450) The Rademacher Complexity of our model family is bounded for all distributions µ∈∆(Rd).

Formal: Model's Rademacher Complexity is bounded for µ∈∆(Rd).

(451) There exists some B > 0 so that for any set of n samples drawn i.i.d.

Formal: Some B > 0 exists for any i.i.d. sample set.

(452) Rµ(Θ) ≤ B√n given that µ is an element of ∆(Rd)

Formal: Rµ(Θ) ≤ B√n if µ∈ ∆(Rd).

(453) Lemma 4 (Rademacher Complexity Generalization Bound): If Assumption 2 holds, then for any θ

Formal: Lemma 4: Under Assumption 2, θ's empirical vs. population error is boundable.

English Reconstruction:

=== English Reconstruction of the Argument ===

- Sharpness-aware minimization (SAM) is a promising method to improve generalization.
- The original proposal of SAM by Foret et al. signifies a foundational shift.
- We study SAM for out-of-distribution (OOD) generalization.
- The original SAM outperforms the Adam baseline by 4.76% in zero-shot out-of-distribution (OOD) ge
- An OOD (out-of-distribution) generalization bound can be provided in terms of sharpness for a particu
- Gradual domain adaptation (GDA) is a form of out-of-distribution (OOD) generalization that utilizes in
- The original SAM outperforms the baseline of Adam on each of the experimental datasets by 0.82%
- The strongest SAM variants outperform Adam by 1.52% on average.
- A generalization bound for SAM is provided in the GDA setting.
- Asymptotically, this generalization bound is no better than the one for self-training in the literature of G
- There is a disconnection between the theoretical justification for SAM and its empirical performance.
- Low sharpness alone does not account for all of SAM’s generalization benefits.
- There are potential avenues for obtaining a tighter analysis for Stochastic Approximation Methods (S
- Theoretical results provide a solid starting point for analyzing SAM in OOD settings.
- SAM can be applied to OOD settings to significantly improve accuracy.
- Newer variants of SAM can be leveraged for further improvements in accuracy.
- Sharpness-Aware Minimization (SAM) is a promising new optimization algorithm.
- A robust optimization procedure can lead to significant performance gains in the i.i.d.

- SAM remains understudied in the out-of-distribution (OOD) generalization setting.
- A number of SAM variants have been proposed to improve the accuracy and efficiency of the origina
- The potential to enhance OOD (out-of-distribution) generalization exists.
- There are eight SAM variants, including the original SAM.
- SAM can be used to improve zero-shot OOD generalization.
- The strongest SAM variants can be used for an even further improvement.
- A theoretical analysis of SAM under the distribution shift setting provides performance gains.
- We extend the setting to gradual domain adaptation.
- SAM outperforms the Adam baseline by 0.82% on average.
- The strongest SAM variants achieve an even greater 1.52% average improvement over Adam.
- SAM and the strongest SAM variants can be used for consistent performance gains in GDA.
- Our work is asymptotically the same as prior work in the GDA literature (Wang et al., 2022).
- We further define the sharpness of θ and the corresponding ρ-robust empirical loss.
- The ρ-robust risk of parameters $\theta \in \Theta$ is the maximum loss obtained by perturbing θ in the worst pos
- The ρ-sharpness of parameters θ measures how much the loss increases when we perturb them in t
- Sharpness-aware minimization (SAM) proposes minimizing the ρ-robust empirical loss rather than th
- SAM’s objective is to find a minimizer θ■ of the form: θ■=argminθ∈Θ max β:■β■2≤ρ E(θ+β).
- The SAM gradient drops a second-order term arising from the chain rule.
- Our results can be extended to any ■p-norm.
- This should not be confused with m-sharpness from Foret et al.
- SAM (Sharpness-Aware Minimization) is a method for optimizing model parameters.
- SAM uses variant-specific oracles for gradient computation, perturbation, and descent steps.
- Adaptive SAM (ASAM) is a version of SAM that uses a scale-invariant version of the first-order appro
- Adaptive SAM has one hyperparameter ρ.
- FisherSAM is a special case of ASAM with a specific normalization operator.
- K-SAM is a variant of SAM that only uses the top K data samples with the highest loss for gradient e
- The idea behind Equation (12) is to remove the component of the descent gradient ∇θE(θ)|θ+β■lying
- Using only the full-batch direction of the ascent-step gradient impairs performance.
- Computing the full-batch gradient $\nabla\theta E(\theta)$ is computationally prohibitive.
- FriendlySAM uses a specific perturbation method defined by an equation.
- NoSAM is a variant of SAM that only performs the SAM perturbation on normalization layers.
- The computational cost is compared across three specific metrics.
- EfficientSAM (ESAM) is a variant of SAM intended to make SAM more efficient.
- Stochastic weight perturbation is applied by performing the SAM perturbation on a fraction of the wei
- Sharpness-sensitive data selection is performed by only computing gradients over data samples with
- A model θS is trained on a source domain $S \in \Delta(X \times Y)$ with a training set.
- The zero-shot OOD generalization error is given by $ET(\theta S)$.
- The intermediate domains are evenly distributed between source and target.
- The intermediate domains are also evenly distributed between source and target.
- Variants such as LookSAM, F-SAM, and FisherSAM offer the strongest and most consistent improve
- LookSAM and NoSAM perform the best among the variants with reduced computational cost.
- LookSAM and NoSAM often outperform the original SAM in addition to being more efficient.
- FisherSAM takes the information geometry of the data into account, using an approximation of the Fi
- Under the cross-entropy loss used specifically in the experiments in Table 2, the Fisher information m
- FisherSAM could be understood as a second-order approximation of the loss perturbation.
- Unlike FisherSAM, the connection between the FriendlySAM objective and OOD performance is not

- The modified perturbation makes FriendlySAM significantly more robust to the choice of the ρ hyperp
- The optimal value of ρ depends intricately on the choice of dataset.
- FriendlySAM provides performance gains in experiments by penalizing sharpness adaptively and sta
- The Wasserstein distance is the smallest cost of moving mass between two distributions measured b
- The loss functions considered in this paper are Lipschitz continuous.
- For the loss function ■, there exist constants $\rho1$, $\rho2$, $\rho3$ such that certain inequalities involving ■ hold
- This result holds for any choice of μ,ν on $Y\times X$.
- Lemma 1 (Sharpness-Aware Error Difference Over Shifted Domains) posits a bound on the differenc
- Population error of a model θ can be bounded in terms of its empirical sharpness.
- For any model $\theta \in \Theta$ satisfying $E(\theta) \leq E■\sim N(0,\rho2I)[E(\theta+■)]$ for some $\rho > 0$, with high probability (w.p
- Using the error difference lemma from Lemma 1 and the PAC-Bayesian bound from Lemma 2, an O
- This result upper bounds the error of a model $\theta\mu$ on domain μ by the error of $\theta\nu$ on domain ν, the sar
- Theorem 1 (Sharpness-Aware Domain Adaptation Error): For distributions μ and ν over $X\times Y$ and an
- Each domain $t\in[T]$ is a distribution μt over $X\times Y$.
- The distribution shift between successive pairs of gradually shifted distributions and the average distr
- In gradual domain adaptation, a learner progressively accesses unlabeled examples from intermedia
- Adding Gaussian perturbation around $\theta\mu$ increases the expected loss.
- We choose the optimal number of intermediate domains T■ for SAM from Figure 1.
- SAM can be applied to GDA to consistently achieve stronger performance.
- SAM is not too sensitive to the perturbation radius hyperparameter.
- SAM with $\rho= 0.2$ leads to the strongest performance on Rotated MNIST, Portraits, and Color MNIST.
- SAM with $\rho= 0.05$ leads to the strongest performance on Covertype.
- The strongest SAM variants lead to an average improvement of 1.42% compared to using Adam.
- The SAM variants with reduced computational cost tend to underperform Adam on GDA.
- FriendlySAM outperforms SAM by 0.40% on average across all datasets.
- FisherSAM slightly underperforms SAM by 0.01%.
- SAM can be used to consistently improve target domain error for GDA.
- The segment claims that there is an extension of Theorem 1 to the setting of gradual domain adaptat
- GDA is presented as a technique which improves target domain error by performing gradual self-trair
- The discrepancy measure captures the non-stationarity of the gradually shifting data.
- The sequential Rademacher complexity generalizes the standard Rademacher complexity to the onli
- Each of the nT samples is viewed as the smallest element of the adaptation process.
- Generalization bound can be stated for GDA performed using SAM.
- The population risk of the gradually adapted model θT can be bounded under certain conditions.
- By applying Corollary 2 of Kuznetsov & Mohri (2020a), a preliminary bound on the error in the target
- The bound we obtain in Theorem 2 is of a specific mathematical form.
- Adding Gaussian perturbation to each solution θt increases the expected loss.
- Theorem 1 must relate the parameters of the successive domains separately.
- The PAC Bayes bound introduces an additional weight norm term W_avg.
- The PAC Bayesian bound yields a different sample complexity term compared to the original analysis
- The overall sample complexity is expected to remain the same between our Theorem 2 and the mair
- A tighter error difference between shifted domains would improve the analysis.
- A localized analysis could exploit implicit properties of SAM.
- There is a relationship between sharpness and generalization.
- Sharpness is among the empirical measures most strongly correlated with generalization.
- Sharp minima can generalize under reparameterizations that cause flat minima to become arbitrarily

- A measure of sharpness is tied to the information geometry of the data and is invariant under reparar
- Many works have explored algorithms that lead to flatter solutions.
- In the context of domain generalization (DG), sharpness affects generalization.
- A modified version of stochastic weight averaging leads to flatter minima with improved DG.
- Generalization bounds depend on the empirical robust loss in the source domain.
- A flatness-aware minimization algorithm for DG leads to improved performance.
- Out-of-distribution (OOD) generalization bounds are presented based on sharpness.
- Sharpness-Aware Minimization (SAM) was originally proposed in Foret et al.
- A PAC Bayesian analysis provides a generalization bound in terms of the expected sharpness over a
- The practical implementation of SAM uses this first-order approximation.
- The flatness of the final solution does not sufficiently capture the generalization benefit from SAM alo
- SAM leads to lower rank features with fewer active ReLU units.
- SAM enhances feature quality by selecting more balanced features.
- SAM enhances robustness to label noise through implicitly regularizing the model Jacobian.
- SAM has an implicit denoising mechanism which prevents harmful overfitting in settings when SGD w
- Many variants of SAM have been proposed to improve the efficiency and accuracy of the original SA
- Gradual self-training (GST) in GDA outperforms standard self-training without intermediate domains.
- These bounds have an exponential dependence on the number of intermediate domains T.
- The analysis can be generalized to any ρ-Lipschitz losses and Wasserstein distances of any order p≥
- The existence of an optimal choice of intermediate domains T is suggested by the refined bounds.
- A new method of generating intermediate domains in an encoded feature space is proposed.
- The main limitation of this work is the discrepancy between our theoretical analysis based on sharpne
- The analysis for SAM can be tightened.
- SAM contributes to out-of-distribution generalization.
- The original SAM achieved a 4.76% average improvement over the Adam baseline.
- The strongest SAM variants achieved an 8.01% average improvement over the Adam baseline.
- An OOD (Out-of-Distribution) generalization bound can be derived based on sharpness.
- We provided an extension of our OOD generalization bound to get a generalization bound based on
- There is a discrepancy between theoretical and empirical results regarding SAM.
- Our theoretical results provide a starting point for doing this.
- Our empirical results suggest that SAM can be used empirically to achieve significant gains for OOD
- Sharpness-aware minimization leads to low-rank features.
- There is a cohesive theory that addresses how learning occurs across different domains.
- Domain generalization can be achieved by seeking flat minima.
- Entropy-sgd biases gradient descent into wide valleys.
- Sharpness-aware minimization generalizes better than SGD.
- Sharp minima can generalize for deep nets.
- Efficient sharpness-aware minimization leads to improved training of neural networks.
- Sharpness-aware minimization efficiently improves generalization.
- Flat minima in the context of optimization and machine learning refer to regions in the parameter spa
- Batch normalization accelerates deep network training.
- Averaging weights leads to wider optima and better generalization.
- Large-batch training for deep learning leads to a generalization gap.
- Large-batch training results in sharp minima.
- Information geometry provides a framework for understanding optimization techniques like sharpness
- Self-training is an effective method for gradual domain adaptation.

- Discrepancy-based theory is useful for forecasting non-stationary time series.
- Adaptive sharpness-aware minimization (Asam) is proposed for scale-invariant learning in deep neur
- The natural gradient method provides new insights and perspectives.
- Normalization layers are all that sharpness-aware minimization needs.
- Online learning can be understood through the framework of sequential complexities.
- Sharpness-aware minimization enhances feature quality via balanced learning.
- Dropout is a simple way to prevent neural networks from overfitting.
- Gradual domain adaptation can be better understood through improved analysis.
- Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization.
- Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time
- A theoretical framework for out-of-distribution generalization is necessary.
- Flatness-aware minimization is a crucial method for improving domain generalization.
- Invariant representations are crucial for effective domain adaptation.
- There are fundamental limits in invariant representation learning.
- Tradeoffs exist in the process of learning invariant representations.
- Gradual domain adaptation via gradient flow is a viable method.
- Robust out-of-distribution generalization can be achieved through considerations of sharpness.
- |Eρμ(θμ)–Eν(θν)|≤Sρ(θμ) +O(■θμ–θν■+Wp(μ,ν))
- Given distributions μ, ν over X × Y and an error function E with loss satisfying Assumption 1 with som
- A sharpness-aware generalization bound, along with Rademacher complexity and robust error differe
- If E(θ) ≤ E■~N(0,ρ2I)[E(θ+■)], then with probability ≥ 1–δ, E(θ) ≤ ˆEρ(θ) + O(sqrt((k ln(■θ■2/ρ2) + ln(
- Theorem 2 is a key claim that is restated within the context of 'Total Sharpness-Aware Error Under G
- The population risk of the gradually adapted model θT can be bounded with high probability.
- The term ET(θT) can be bounded by a sequence of steps applying Theorem 1.
- ET–1(θT) and other successive terms (ET–2(θT), ET–3(θT), etc.) can be bound similarly using the de
- The text provides a bound on the expected value ET(θT) given initial conditions and average values o
- |Eμ(θ)–Eν(θ)|≤O (Wp(μ,ν))
- Proposition 1 (Discrepancy Bound - Lemma 2 of Wang et al.
- disc(qt) ≤ O/parenleft| iggt–1/summationdisplay k=0 qk(t–k–1)Wp(μk,μk+1)/parenright| igg
- disc(qt) ≤ O(tΔ) when qt=q■ t:= (1/t,..., 1/t)
- Definition 8 (Rademacher Complexity) introduces the concept of empirical Rademacher complexity fo
- The Rademacher Complexity of our model family is bounded for all distributions μ∈Δ(Rd).
- There exists some B > 0 so that for any set of n samples drawn i.i.d.
- Rμ(Θ) ≤ B√n given that μ is an element of Δ(Rd)
- Lemma 4 (Rademacher Complexity Generalization Bound): If Assumption 2 holds, then for any θ∈Θ,