

A Appendix

A.1 Relevance to NeurIPS

The ML community is deeply motivated by a desire to have a positive impact on the world. This desire is reflected in recent efforts in the ML community, such as NeurIPS’s requirement for the inclusion of broader impacts statements for all submitted papers in 2020, the Resistance AI Workshop at NeurIPS 2020 which investigated how AI concentrates power, and the Navigating the Broader Impacts of AI Research at NeurIPS 2020 which sought to understand the impacts of ML research as a whole on society. Understanding what the social impact of a paper, let alone the discipline, is difficult. Merely looking at various benchmarks or broader impact statements, for example, is insufficient. This paper attempts to begin to bridge this gap by seeking to understand the value commitments in papers published at NeurIPS and a closely related conference, ICML. As such, this paper is highly relevant to the NeurIPS audience. While research into core technical ML topics – reinforcement learning, deep learning, optimization, etc. – are vital to NeurIPS and the wider ML community, so is research on where these research areas stand with regard to societal impact, both in a positive and negative manner, as well as the benefits they bring and to whom.

A.2 Additional Methodological Details

A.2.1 Data Sources

To determine the most-cited papers from each conference, we rely on the publicly-available Semantic Scholar database, which includes bibliographic information for scientific papers, including citation counts.⁶ Using this data, we chose the most cited papers from each of 2008, 2009, 2018, 2019 published at NeurIPS and ICML.

Like all bibliographic databases, Semantic Scholar is imperfect, and thus our selection includes one paper that was actually published in 2010, and one that was retracted from NeurIPS prior to publication (see §A.8 for details). In addition, the citations counts used to determine the most cited papers reflect a static moment in time, and may differ from other sources.

Because all data used for this paper (aside from the actual annotations, which we contribute) have been previously published at NeurIPS or ICML, we chose not to seek permission to annotate this data from the original authors. Similarly, although it is possible that the original papers may contain personally identifying information or offensive content, we rely on the fact that the original authors contributed their work to the same community to which our own work is directed, and we thus believe that the potential harm from this is minimal.

A.2.2 Defining elite university

To determine the list of elite universities, we follow Ahmed and Wahed [4], and rely on the QS World University Rankings for the discipline of computer science. For 2018/19, we take the top 50 schools from the CS rankings for 2018. For 2008/09, we take the top 50 schools from the CS rankings for 2011, as the closest year for which data is available.

A.2.3 Defining big tech

We used Abdalla and Abdalla’s [2] criterion to what is considered "big tech", which is comprised of: Alibaba, Amazon, Apple, Element AI, Facebook, Google, Huawei, IBM, Intel, Microsoft, Nvidia, Open AI, Samsung, and Uber. Furthermore, we added DeepMind to this list. We considered all other companies as "non-big tech."

A.3 Annotations

We include the annotations of all papers in the supplementary zip file. To present a birds-eye view of the value annotations, we present randomly selected examples of annotated sentences in section §A.7. In addition, we present the frequency of occurrence for all values (prior to grouping) in Figure A.1 below.

⁶<http://s2-public-api.prod.s2.allenai.org/corpus/>

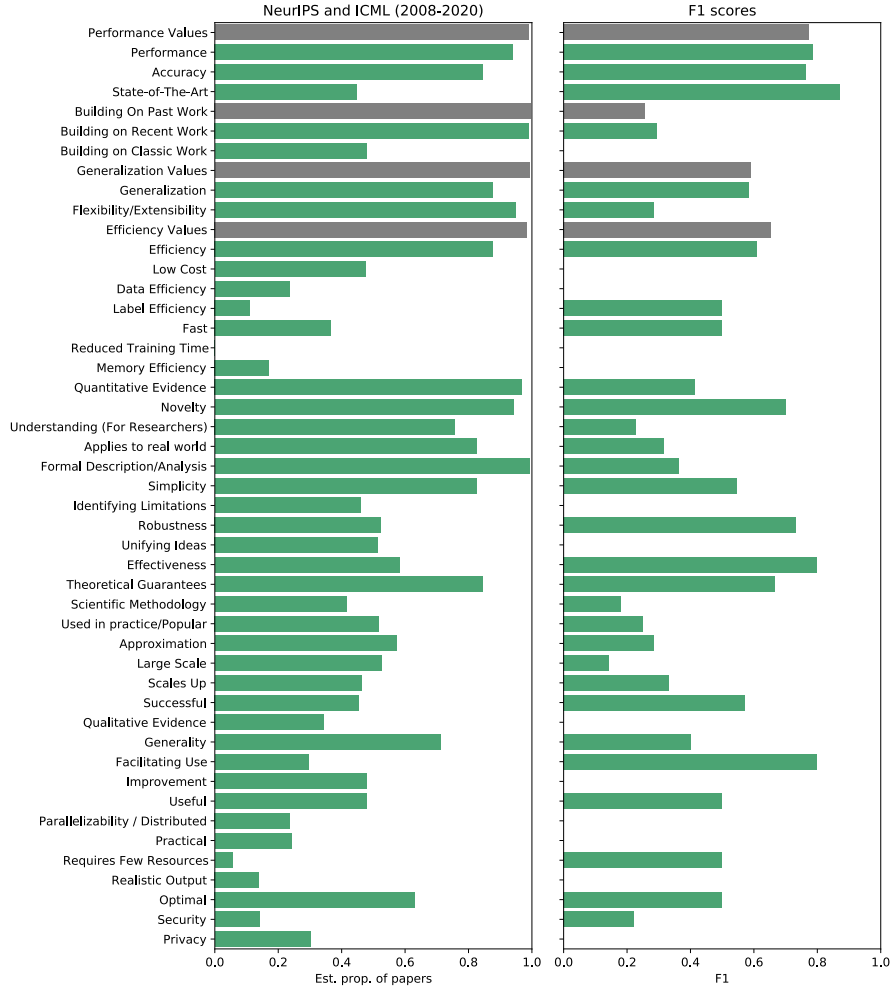


Figure A.2: Proportion of papers in from 2008–2020 (combining NeurIPS and ICML) predicted to have at least one sentence expressing each value (left), and estimated performance (F1) of the corresponding classifiers (right). Note that the overall performance of most classifiers is generally poor, indicating that the estimates on the left should be treated as unreliable in most cases. Grey bars represent the grouped values. Classifier were not trained for values with less than 20 representative sentences.

We then apply the classifiers trained above to each sentence in each paper. For each value, we then compute the proportion of papers (combining NeurIPS and ICML for this entire time period) that had at least one sentence predicted to exhibit that value. The overall proportions are shown in Figure A.2 (left). As can be seen, the relative prevalence of values is broadly similar to our annotated sample, though many are predicted to occur with greater frequency, as expected. However, to reiterate, we should be highly skeptical of these findings, given the poor performance of the classifiers.

Finally, as an additional exploration, we focus on the Performance-related values (*Performance*, *Accuracy*, and *State-of-the-art*), which represent the overall most prevalent group in our annotations, and were relatively easy to identify using classification, and plot the estimated frequency over time for both conferences (For NeurIPS, which has better archival practices, we extend the analysis back to 1987). We should again treat these results with caution, given all the caveats above, as well as the fact that we are now applying these classifiers outside the temporal range from which the annotations were collected. Nevertheless, the results, shown in Figure A.3 suggest that these values have gradually become more common in NeurIPS over time, reinforcing the contingent nature of the dominance of the current set of values. Further investigation is required, however, in order to verify this finding.

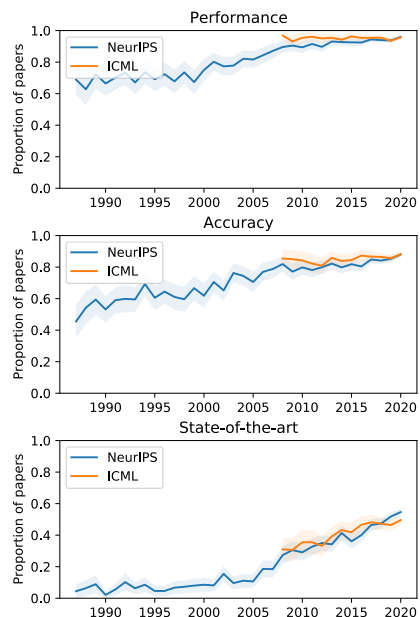


Figure A.3: Proportion of papers per year (of those published in ICML and NeurIPS) that are classified as having at least one sentence expressing *Performance*, *Accuracy*, or *State-of-the-art*, (top, middle, and bottom), based on simple text classifiers trained on our annotations. Bands show ± 2 standard deviations, reflecting the changing overall number of papers per year.

A.5 Code and Reproducibility

We include, in the supplementary zip file, the code used for all data analysis; in particular, we include the code used to run the text classification experiments and generate all figures in the paper. The text classification experiments were run on a 2019 Macbook Air.

A.6 Potential Negative Societal Impacts

Because this paper relies only on manual annotation of papers already published at NeurIPS and ICML, we believe that the potential negative societal impacts of carrying out these annotations and sharing them are minimal. However, we still briefly comment on this here.

First, in terms of the specific concerns highlighted in the NeurIPS call for papers, we believe our annotation work poses no risk to living beings, human rights concerns, threats to livelihoods, etc. Similarly, all annotators are co-authors on this paper, thus there was no risk to participants, beyond what we chose to take on for ourselves.

One area of potential concern might be in terms of unintentionally casting certain authors in a negative light, or unintentionally contributing to harmful tensions within the ML community. In order to minimize the risk of the former, we have chosen to include randomly selected examples but omit author attributions from quoted sources in the main paper. However, we do include a full list of cited papers below, so as to both acknowledge this work, but not draw attention to any one particular source.

Although our intention is to broaden the conversation, we do acknowledge that some authors may perceive our work as being not representative of the type of work they would like to see at NeurIPS, and possibly detrimental to the conference. However, because of the prominence of machine learning today, we feel it is especially important to have these conversations at the premier venues, and hope that our paper will be the basis for useful conversations and future work.

A.7 Random Examples

The list below contains 100 random examples drawn from the annotated data, along with the set of annotated values for each. These sentences were annotated for values within the context of the entire paper.

- The problem of minimizing the rank of a matrix variable subject to certain constraints arises in many fields including machine learning, automatic control, and image compression. **Used in practice/Popular**
- Locality-sensitive hashing [6] is an effective technique that performs approximate nearest neighbor searches in time that is sub-linear in the size of the database **Approximation, Building on recent work, Effectiveness, Fast**
- In the finite case, analysis of optimization and generalization of fully-trained nets is of course an open problem **Formal description/analysis, Generalization**
- So to achieve adversarial robustness, a classifier must generalize in a stronger sense. **Generalization, Robustness**
- Robustness to label corruption is similarly improved by wide margins, such that pre-training alone outperforms certain task-specific methods, sometimes even after combining these methods with pre-training. **Performance, Robustness, Understanding (for researchers)**
- RBMs have been particularly successful in classification problems either as feature extractors for text and image data (Gehler et al., 2006) or as a good initial training phase for deep neural network classifiers (Hinton, 2007). **Building on recent work, Flexibility/Extensibility, Successful**
- Our theoretical analysis naturally leads to a new formulation of adversarial defense which has several appealing properties; in particular, it inherits the benefits of scalability to large datasets exhibited by Tiny ImageNet, and the algorithm achieves state-of-the-art performance on a range of benchmarks while providing theoretical guarantees. **Robustness, Scales up, Security, Theoretical guarantees**
- The current paper focuses on the training loss, but does not address the test loss. **Generalization**
- This result is significant since stochastic methods are highly preferred for their efficiency over deterministic gradient methods in machine learning applications. **Efficiency**
- Ranking, which is to sort objects based on certain factors, is the central problem of applications such as information retrieval (IR) and information filtering. **Applies to real world, Used in practice/Popular**
- This subspace is important, because, when projected onto this subspace, the means of the distributions are well-separated, yet the typical distance between points from the same distribution is smaller than in the original space. **Important**
- Overall, the existence of such adversarial examples raises concerns about the robustness of current classifiers. **Identifying limitations, Robustness**
- We have shown that biased compressors if naively used can lead to bad generalization, and even non-convergence. **Formal description/analysis, Generalization**
- Bartlett and Mendelson [2002] provide a generalization bound for Lipschitz loss functions. **Building on classic work, Generalization**
- The principal advantage of taking this “lateral” approach arises from the fact that compact representation in trajectory space is better motivated physically than compact representation in shape space **Realistic world model**
- In this paper, we show that gradient descent on deep overparametrized networks can obtain zero training loss **Formal description/analysis, Theoretical guarantees**
- Moreover, web queries often have different meanings for different users (a canonical example is the query jaguar) suggesting that a ranking with diverse documents may be preferable. **Diverse output, User influence**

- We include human performance estimates for all benchmark tasks, which verify that substantial headroom exists between a strong BERT-based baseline and human performance. **Learning from humans, Performance**
- In this paper we propose a simple and fast algorithm SVP (Singular Value Projection) for rank minimization under affine constraints (ARMP) and show that SVP recovers the minimum rank solution for affine constraints that satisfy a restricted isometry property (RIP). **Fast, Novelty, Simplicity**
- We use standard formalization of multiclass classification, where data consists of sample x and its label y (an integer from 1 to k). **Building on classic work**
- A number of recent works has shown that the low rank solution can be recovered exactly via minimizing the trace norm under certain conditions (Recht et al., 2008a; Recht et al., 2008b; Candes Recht, 2008). **Building on recent work**
- This difficulty has necessitated the use of a heuristic inference procedure, that nonetheless was accurate enough for successful learning. **Accuracy, Successful**
- We illustrate such potential by measuring search space properties relevant to architecture search. **Quantitative evidence (e.g. experiments)**
- Deep architectures consist of feature detector units arranged in layers. Lower layers detect simple features and feed into higher layers, which in turn detect more complex features. **Simplicity**
- This makes the updates hard to massively parallelize at a coarse, data-parallel level (e.g., by computing the updates in parallel and summing them together centrally) without losing the critical stochastic nature of the updates. **Large scale, Parallelizability / distributed**
- This suggests future work on model robustness should evaluate proposed methods with pretraining in order to correctly gauge their utility, and some work could specialize pretraining for these downstream tasks. **Robustness**
- Adversarial training remains among the most trusted defenses, but it is nearly intractable on large scale problems. **Scales up, Security**
- For complex robots such as humanoids or light-weight arms, it is often hard to model the system sufficiently well and, thus, modern regression methods offer a viable alternative [7,8]. **Realistic world model**
- In contrast to prior work that operates in this goal-setting model, we use states as goals directly, which allows for simple and fast training of the lower layer. **Reduced training time, Simplicity**
- Meanwhile, using less resources tends to produce less compelling results (Negrinho Gordon, 2017; Baker et al., 2017a). **Requires few resources**
- This finding represents an exciting opportunity for defense against neural fake news: the best models for generating neural disinformation are also the best models at detecting it. **Applies to real world**
- Our strong empirical results suggest that randomized smoothing is a promising direction for future research into adversarially robust classification. **Quantitative evidence (e.g. experiments), Robustness, Security**
- We then turn our attention to identifying the roots of BatchNorm’s success. **Successful, Understanding (for researchers)**
- We also report the results of large-scale experiments comparing these three methods which demonstrate the benefits of the mixture weight method: this method consumes less resources, while achieving a performance comparable to that of standard approaches. **Large scale, Performance, Requires few resources**
- This paper does not cover the generalization of over-parameterized neural networks to the test data. **Avoiding train/test discrepancy, Generalization**
- While there has been success with robust classifiers on simple datasets [31, 36, 44, 48], more complicated datasets still exhibit a large gap between “standard” and robust accuracy [3, 11]. **Applies to real world, Robustness, Successful**

- In this paper, we have shown theoretically how independence between examples can make the actual effect much smaller. **Novelty, Theoretical guarantees**
- We provide empirical evidence that several recently-used methods for estimating the probability of held-out documents are inaccurate and can change the results of model comparison. **Accuracy, Building on recent work, Quantitative evidence (e.g. experiments)**
- This agreement is robust across different architectures, optimization methods, and loss functions **Robustness**
- Unfortunately, due to the slow-changing policy in an actor-critic setting, the current and target value estimates remain too similar to avoid maximization bias. **Accuracy**
- As a future work, we are pursuing a better understanding of probabilistic distributions on the Grassmann manifold. **Understanding (for researchers)**
- We also view these results as an opportunity to encourage the community to pursue a more systematic investigation of the algorithmic toolkit of deep learning and the underpinnings of its effectiveness. **Effectiveness, Understanding (for researchers)**
- This challenge is further exacerbated in continuous state and action spaces, where a separate actor network is often used to perform the maximization in Q-learning. **Performance**
- The vulnerability of neural networks to adversarial perturbations has recently been a source of much discussion and is still poorly understood. **Robustness, Understanding (for researchers)**
- Most of the evaluation methods described in this paper extend readily to more complicated topic models— including non-parametric versions based on hierarchical Dirichlet processes (Teh et al., 2006)—since they only require a MCMC algorithm for sampling the latent topic assignments z for each document and a way of evaluating probability $P(w | z, m)$. **Flexibility/Extensibility, Understanding (for researchers)**
- In a formulation closely related to the dual problem, we have: $w^* = \operatorname{argmin}_w F(w) - c \sum_{i=1}^n \langle w, x_i \rangle$ where, instead of regularizing, a hard restriction over the parameter space is imposed (by the constant c). **Formal description/analysis**
- Second, we evaluate a surrogate loss function from four aspects: (a) consistency, (b) soundness, (c) mathematical properties of continuity, differentiability, and convexity, and (d) computational efficiency in learning. **Efficiency**
- This leads to two natural questions that we try to answer in this paper: (1) Is it feasible to perform optimization in this very large feature space with cost which is polynomial in the size of the input space? **Performance**
- Despite its pervasiveness, the exact reasons for BatchNorm’s effectiveness are still poorly understood. **Understanding (for researchers)**
- We have presented confidence-weighted linear classifiers, a new learning method designed for NLP problems based on the notion of parameter confidence. **Novelty**
- In addition, the experiments reported here suggest that (like other strategies recently proposed to train deep deterministic or stochastic neural networks) the curriculum strategies appear on the surface to operate like a regularizer, i.e., their beneficial effect is most pronounced on the test set. **Beneficence, Quantitative evidence (e.g. experiments)**
- These give further insight into hash-spaces and explain previously made empirical observations. **Understanding (for researchers)**
- This means that current algorithms reach their limit at problems of size 1TB whenever the algorithm is I/O bound (this amounts to a training time of 3 hours), or even smaller problems whenever the model parametrization makes the algorithm CPU bound. **Memory efficiency, Reduced training time**
- Much of the results presented were based on the assumption that the target distribution is some mixture of the source distributions. **Valid assumptions**
- Empirical investigation revealed that this agrees well with actual training dynamics and predictive distributions across fully-connected, convolutional, and even wide residual network architectures, as well as with different optimizers (gradient descent, momentum, mini-batching) and loss functions (MSE, cross-entropy). **Generalization, Quantitative evidence (e.g. experiments), Understanding (for researchers)**

- We design a new spectral norm that encodes this a priori assumption, without the prior knowledge of the partition of tasks into groups, resulting in a new convex optimization formulation for multi-task learning. **Novelty**
- Recent progress in natural language generation has raised dual-use concerns. **Progress**
- These kernel functions can be used in shallow architectures, such as support vector machines (SVMs), or in deep kernel-based architectures that we call multilayer kernel machines (MKMs). **Flexibility/Extensibility**
- Using MCMC instead of variational methods for approximate inference in Bayesian matrix factorization models leads to much larger improvements over the MAP trained models, which suggests that the assumptions made by the variational methods about the structure of the posterior are not entirely reasonable. **Understanding (for researchers)**
- In particular, the deep belief network (DBN) (Hinton et al., 2006) is a multilayer generative model where each layer encodes statistical dependencies among the units in the layer below it; it is trained to (approximately) maximize the likelihood of its training data. **Approximation, Data efficiency**
- Furthermore, the learning accuracy and performance of our LGP approach will be compared with other important standard methods in Section 4, e.g., LWPR [8], standard GPR [1], sparse online Gaussian process regression (OGP) [5] and -support vector regression (-SVR) [11], respectively **Accuracy, Performance, Quantitative evidence (e.g. experiments)**
- • propose a simple method based on weighted minibatches to stochastically train with arbitrary weights on the terms of our decomposition without any additional hyperparameters. **Efficiency, Simplicity**
- For example, Ng (2004) examined the task of PAC learning a sparse predictor and analyzed cases in which an 1 constraint results in better solutions than an 2 constraint. **Building on recent work**
- Graph Convolutional Networks (GCNs) (Kipf Welling, 2017) are an efficient variant of Convolutional Neural Networks (CNNs) on graphs. GCNs stack layers of learned first-order spectral filters followed by a nonlinear activation function to learn graph representations. **Efficiency**
- This is a linear convergence rate. **Building on recent work, Efficiency, Quantitative evidence (e.g. experiments), Theoretical guarantees**
- However, as we observe more interactions, this could emerge as a clear feature. **Building on recent work, Data efficiency**
- Here we propose the first method that supports arbitrary low accuracy and even biased compression operators, such as in (Alistarh et al., 2018; Lin et al., 2018; Stich et al., 2018). **Accuracy, Novelty**
- Much recent work has been done on understanding under what conditions we can learn a mixture model. **Understanding (for researchers)**
- For this reason, we present an extension of the standard greedy OMP algorithm that can be applied to general structured sparsity problems, and more importantly, meaningful sparse recovery bounds can be obtained for this algorithm. **Building on recent work**
- In this paper we show that this assumption is indeed necessary: by considering a simple yet prototypical example of GAN training we analytically show that (unregularized) GAN training is not always locally convergent **Formal description/analysis**
- Overestimation bias is a property of Q-learning in which the maximization of a noisy value estimate induces a consistent overestimation **Accuracy**
- This drawback prevents GPR from applications which need large amounts of training data and require fast computation, e.g., online learning of inverse dynamics model for model-based robot control **Fast, Large scale**
- This is problematic since we find there are techniques which do not comport well with pre-training; thus some evaluations of robustness are less representative of real-world performance than previously thought. **Applies to real world, Performance, Robustness**

- Approximation of this prior structure through simple, efficient hyperparameter optimization steps is sufficient to achieve these performance gains **Approximation, Efficiency, Performance, Simplicity**
- The second mysterious phenomenon in training deep neural networks is “deeper networks are harder to train.” **Performance**
- However, the definition of our metric is sufficiently general that it could easily be used to test, for example, invariance of auditory features to rate of speech, or invariance of textual features to author identity. **Generalization**
- In Sec. 6 we test the proposed algorithm for face recognition and object categorization tasks. **Applies to real world, Quantitative evidence (e.g. experiments)**
- It is possible to train classification RBMs directly for classification performance; the gradient is fairly simple and certainly tractable. **Performance**
- Figure 1 contrasts these two approaches. Defining and evaluating models using ODE solvers has several benefits: **Beneficence**
- They claim to achieve 12×2 radius of 3 (for images with pixels in $[0, 1]$). **Generalization, Robustness**
- Two commonly used penalties are the 1- norm and the square of the 2-norm of w . **Used in practice/Popular**
- What should platforms do? Video-sharing platforms like YouTube use deep neural networks to scan videos while they are uploaded, to filter out content like pornography (Hosseini et al., 2017). **Applies to real world**
- We mention various properties of this penalty, and provide conditions for the consistency of support estimation in the regression setting. Finally, we report promising results on both simulated and real data **Applies to real world**
- There could be a separate feature for “high school student,” “male,” “athlete,” and “musician” and the presence or absence of each of these features is what defines each person and determines their relationships. **Building on recent work**
- So, the over-parameterized convergence theory of DNN is much simpler than that of RNN. **Simplicity, Understanding (for researchers)**
- Other threat models are possible: for instance, an adversary might generate comments or have entire dialogue agents, they might start with a human-written news article and modify a few sentences, and they might fabricate images or video. **Learning from humans**
- More generally, we hope that future work will be able to avoid relying on obfuscated gradients (and other methods that only prevent gradient descent-based attacks) for perceived robustness, and use our evaluation approach to detect when this occurs. **Generality, Robustness**
- For example, the learned linear combination does not consistently outperform either the uniform combination of base kernels or simply the best single base kernel (see, for example, UCI dataset experiments in [9, 12], see also NIPS 2008 workshop). **Performance**
- Our main contributions are:
 - We analyze GP-UCB, an intuitive algorithm for GP optimization, when the function is either sampled from a known GP, or has low RKHS norm. **Optimal**
- For the standard linear setting, Dani et al. (2008) provide a near-complete characterization explicitly dependent on the dimensionality. In the GP setting, the challenge is to characterize complexity in a different manner, through properties of the kernel function. **Building on classic work**
- This allows us to map each architecture A to its approximate hyperparameter-optimized accuracy **Accuracy**
- Unfortunately, they could only apply their method to linear networks. **Flexibility/Extensibility**
- The strength of the adversary then allows for a trade-off between the enforced prior, and the data-dependent features. **Understanding (for researchers)**

- We observe that the computational bottleneck of NAS is the training of each child model to convergence, only to measure its accuracy whilst throwing away all the trained weights. **Accuracy**
- We show that the number of subproblems need only be logarithmic in the total number of possible labels, making this approach radically more efficient than others. **Efficiency**
- We establish a new notion of quadratic approximation of the neural network, and connect it to the SGD theory of escaping saddle points. **Novelty, Unifying ideas or integrating components**
- In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors. **Accuracy, Robustness, Theoretical guarantees**
- A limit on the number of queries can be a result of limits on other resources, such as a time limit if inference time is a bottleneck or a monetary limit if the attacker incurs a cost for each query. **Applies to real world, Low cost, Requires few resources**
- Preliminary experiments demonstrate that it is significantly faster than batch alternatives on large datasets that may contain millions of training examples, yet it does not require learning rate tuning like regular stochastic gradient descent methods. **Quantitative evidence (e.g. experiments), Reduced training time**
- SuperGLUE is available at super.gluebenchmark.com. **Facilitating use (e.g. sharing code)**

A.8 Full List of Cited Papers

The full list of annotated papers is given below, along with the annotated scores (in square brackets) for *Discussion of Negative Potential* [left] (0 = Doesn't mention negative potential; 1 = Mentions but does not discuss negative potential; 2 = Discusses negative potential) and *Justification* [right] (0 = Doesn't rigorously justify how it achieves technical goal; 1 = Justifies how it achieves technical goal but no mention of societal need; 2 = States but does not justify how it connects to a societal need; 3 = States and somewhat justifies how it connects to a societal need; 4 = States and rigorously justifies how it connects to a societal need). Note that due to minor errors in the data sources used, the distribution of papers over venues and years is not perfectly balanced. For the same reason, the list also contains one paper from 2010 (rather than 2009), as well as one paper that was retracted before publication at NeurIPS (marked with a *).

- Mingxing Tan, Quoc Le. [EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks](#). In *Proceedings of ICML*, 2019. [0/1]
- Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, Ruosong Wang. [Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks](#). In *Proceedings of ICML*, 2019. [0/1]
- Jeremy Cohen, Elan Rosenfeld, Zico Kolter. [Certified Adversarial Robustness via Randomized Smoothing](#). In *Proceedings of ICML*, 2019. [0/1]
- Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, Michael Jordan. [Theoretically Principled Trade-off between Robustness and Accuracy](#). In *Proceedings of ICML*, 2019. [0/2]
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu. [MASS: Masked Sequence to Sequence Pre-training for Language Generation](#). In *Proceedings of ICML*, 2019. [0/1]
- Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, Kilian Weinberger. [Simplifying Graph Convolutional Networks](#). In *Proceedings of ICML*, 2019. [0/1]
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, Vaishal Shankar. [Do ImageNet Classifiers Generalize to ImageNet?](#) In *Proceedings of ICML*, 2019. [0/2]
- Justin Gilmer, Nicolas Ford, Nicholas Carlini, Ekin Cubuk. [Adversarial Examples Are a Natural Consequence of Test Error in Noise](#). In *Proceedings of ICML*, 2019. [0/1]

- 912 • Chris Ying, Aaron Klein, Eric Christiansen, Esteban Real, Kevin Murphy, Frank Hutter.
913 [NAS-Bench-101: Towards Reproducible Neural Architecture Search.](#) In *Proceedings of*
914 *ICML*, 2019. [0/2]
- 915 • Dan Hendrycks, Kimin Lee, Mantas Mazeika. [Using Pre-Training Can Improve Model](#)
916 [Robustness and Uncertainty.](#) In *Proceedings of ICML*, 2019. [0/1]
- 917 • Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, Martin Jaggi. [Error Feedback](#)
918 [Fixes SignSGD and other Gradient Compression Schemes.](#) In *Proceedings of ICML*, 2019.
919 [0/1]
- 920 • Anastasia Koloskova, Sebastian Stich, Martin Jaggi. [Decentralized Stochastic Optimization](#)
921 [and Gossip Algorithms with Compressed Communication.](#) In *Proceedings of ICML*, 2019.
922 [0/2]
- 923 • Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena. [Self-Attention Generative](#)
924 [Adversarial Networks.](#) In *Proceedings of ICML*, 2019. [0/1]
- 925 • Zeyuan Allen-Zhu, Yuanzhi Li, Zhao Song. [A Convergence Theory for Deep Learning via](#)
926 [Over-Parameterization.](#) In *Proceedings of ICML*, 2019. [0/1]
- 927 • Simon Du, Jason Lee, Haochuan Li, Liwei Wang, Xiyu Zhai. [Gradient Descent Finds Global](#)
928 [Minima of Deep Neural Networks.](#) In *Proceedings of ICML*, 2019. [0/1]
- 929 • Anish Athalye, Nicholas Carlini, David Wagner. [Obfuscated Gradients Give a False Sense](#)
930 [of Security: Circumventing Defenses to Adversarial Examples.](#) In *Proceedings of ICML*,
931 2018. [0/2]
- 932 • Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, Jeff Dean. [Efficient Neural Architecture](#)
933 [Search via Parameters Sharing.](#) In *Proceedings of ICML*, 2018. [0/1]
- 934 • Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine. [Soft Actor-Critic: Off-Policy](#)
935 [Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor.](#) In *Proceedings*
936 *of ICML*, 2018. [0/2]
- 937 • Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam
938 Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, Koray Kavukcuoglu. [IMPALA:](#)
939 [Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures.](#) In
940 *Proceedings of ICML*, 2018. [0/1]
- 941 • Scott Fujimoto, Herke Hoof, David Meger. [Addressing Function Approximation Error in](#)
942 [Actor-Critic Methods.](#) In *Proceedings of ICML*, 2018. [0/1]
- 943 • Hyunjik Kim, Andriy Mnih. [Disentangling by Factorising.](#) In *Proceedings of ICML*, 2018.
944 [0/0]
- 945 • Lars Mescheder, Andreas Geiger, Sebastian Nowozin. [Which Training Methods for GANs](#)
946 [do actually Converge?](#) In *Proceedings of ICML*, 2018. [0/1]
- 947 • Sanjeev Arora, Rong Ge, Behnam Neyshabur, Yi Zhang. [Stronger generalization bounds for](#)
948 [deep nets via a compression approach.](#) In *Proceedings of ICML*, 2018. [0/3]
- 949 • Andrew Ilyas, Logan Engstrom, Anish Athalye, Jessy Lin. [Black-box Adversarial Attacks](#)
950 [with Limited Queries and Information.](#) In *Proceedings of ICML*, 2018. [0/2]
- 951 • Niranjan Srinivas, Andreas Krause, Sham Kakade, Matthias Seeger. [Gaussian Process](#)
952 [Optimization in the Bandit Setting: No Regret and Experimental Design.](#) In *Proceedings of*
953 *ICML*, 2010. [0/1]
- 954 • Honglak Lee, Roger Grosse, Rajesh Ranganath and Andrew Ng. [Convolutional deep belief](#)
955 [networks for scalable unsupervised learning of hierarchical representations.](#) In *Proceedings*
956 *of ICML*, 2009. [0/1]
- 957 • Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro. [Online dictionary learning](#)
958 [for sparse coding.](#) In *Proceedings of ICML*, 2009. [0/1]
- 959 • Yoshua Bengio, Jerome Louradour, Ronan Collobert and Jason Weston. [Curriculum learning.](#)
960 In *Proceedings of ICML*, 2009. [0/1]
- 961 • Laurent Jacob, Guillaume Obozinski and Jean-Philippe Vert. [Group Lasso with Overlaps](#)
962 [and Graph Lasso.](#) In *Proceedings of ICML*, 2009. [0/3]

- 963 • Chun-Nam Yu and Thorsten Joachims. [Learning structural SVMs with latent variables.](#) In
964 *Proceedings of ICML*, 2009. [0/2]
- 965 • Kilian Weinberger, Anirban Dasgupta, Josh Attenberg, John Langford and Alex Smola.
966 [Feature hashing for large scale multitask learning.](#) In *Proceedings of ICML*, 2009. [0/2]
- 967 • Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. [Evaluation methods](#)
968 [for topic models.](#) In *Proceedings of ICML*, 2009. [0/1]
- 969 • Kamalika Chaudhuri, Sham Kakade, Karen Livescu and Karthik Sridharan. [Multi-view](#)
970 [clustering via canonical correlation analysis.](#) In *Proceedings of ICML*, 2009. [0/2]
- 971 • Shuiwang Ji and Jieping Ye. [An accelerated gradient method for trace norm minimization.](#)
972 In *Proceedings of ICML*, 2009. [0/3]
- 973 • Junzhou Huang, Tong Zhang and Dimitris Metaxas. [Learning with structured sparsity.](#) In
974 *Proceedings of ICML*, 2009. [0/1]
- 975 • Rajat Raina, Anand Madhavan and Andrew Ng. [Large-scale deep unsupervised learning](#)
976 [using graphics processors.](#) In *Proceedings of ICML*, 2009. [0/2]
- 977 • Ronan Collobert and Jason Weston. [A unified architecture for natural language processing:](#)
978 [deep neural networks with multitask learning.](#) In *Proceedings of ICML*, 2008. [0/2]
- 979 • Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. [Extracting](#)
980 [and composing robust features with denoising autoencoders.](#) In *Proceedings of ICML*, 2008.
981 [0/1]
- 982 • Ruslan Salakhutdinov and Andriy Mnih. [Bayesian probabilistic matrix factorization using](#)
983 [Markov chain Monte Carlo.](#) In *Proceedings of ICML*, 2008. [0/1]
- 984 • John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. [Efficient projections](#)
985 [onto the \$l_1\$ -ball for learning in high dimensions.](#) In *Proceedings of ICML*, 2008. [0/1]
- 986 • Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathya Keerthi, and S. Sundararajan. [A](#)
987 [dual coordinate descent method for large-scale linear SVM.](#) In *Proceedings of ICML*, 2008.
988 [0/1]
- 989 • Tijmen Tieleman. [Training restricted Boltzmann machines using approximations to the](#)
990 [likelihood gradient.](#) In *Proceedings of ICML*, 2008. [0/1]
- 991 • Hugo Larochelle and Yoshua Bengio. [Classification using discriminative restricted Boltz-](#)
992 [mann machines.](#) In *Proceedings of ICML*, 2008. [0/1]
- 993 • Jihun Hamm and Daniel Lee. [Grassmann discriminant analysis: a unifying view on subspace-](#)
994 [based learning.](#) In *Proceedings of ICML*, 2008. [0/1]
- 995 • Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. [Listwise Approach to](#)
996 [Learning to Rank - Theory and Algorithm.](#) In *Proceedings of ICML*, 2008. [0/1]
- 997 • Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. [Learning diverse rankings with](#)
998 [multi-armed bandits.](#) In *Proceedings of ICML*, 2008. [0/1]
- 999 • Mark Dredze, Koby Crammer, and Fernando Pereira. [Confidence-weighted linear classifica-](#)
1000 [tion.](#) In *Proceedings of ICML*, 2008. [0/1]
- 1001 • Ruslan Salakhutdinov and Iain Murray. [On the quantitative analysis of deep belief networks.](#)
1002 In *Proceedings of ICML*, 2008. [0/1]
- 1003 • Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, Quoc
1004 V. Le. [XLNet: Generalized Autoregressive Pretraining for Language Understanding.](#) In
1005 *Proceedings of NeurIPS*, 2019. [0/1]
- 1006 • Alexis CONNEAU, Guillaume Lample. [Cross-lingual Language Model Pretraining.](#) In
1007 *Proceedings of NeurIPS*, 2019. [0/4]
- 1008 • Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Alek-
1009 sander Madry. [Adversarial Examples Are Not Bugs, They Are Features.](#) In *Proceedings of*
1010 *NeurIPS*, 2019. [0/1]
- 1011 • Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha
1012 Sohl-Dickstein, Jeffrey Pennington. [Wide Neural Networks of Any Depth Evolve as Linear](#)
1013 [Models Under Gradient Descent.](#) In *Proceedings of NeurIPS*, 2019. [0/1]

- 1014 • David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, Colin
1015 A. Raffel. [MixMatch: A Holistic Approach to Semi-Supervised Learning](#). In *Proceedings*
1016 *of NeurIPS*, 2019. [0/1]
- 1017 • Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan,
1018 Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas
1019 Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy,
1020 Benoit Steiner, Lu Fang, Junjie Bai, Soumith Chintala. [PyTorch: An Imperative Style,](#)
1021 [High-Performance Deep Learning Library](#). In *Proceedings of NeurIPS*, 2019. [0/1]
- 1022 • Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Russ R. Salakhutdinov, Ruosong Wang.
1023 [On Exact Computation with an Infinitely Wide Neural Net](#). In *Proceedings of NeurIPS*,
1024 2019. [0/1]
- 1025 • Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao,
1026 Ming Zhou, Hsiao-Wuen Hon. [Unified Language Model Pre-training for Natural Language](#)
1027 [Understanding and Generation](#). In *Proceedings of NeurIPS*, 2019. [0/1]
- 1028 • Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph
1029 Studer, Larry S. Davis, Gavin Taylor, Tom Goldstein. [Adversarial Training for Free!](#) In
1030 *Proceedings of NeurIPS*, 2019. [0/3]
- 1031 • Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee. [ViLBERT: Pretraining Task-Agnostic](#)
1032 [Visiolinguistic Representations for Vision-and-Language Tasks](#). In *Proceedings of NeurIPS*,
1033 2019. [0/1]
- 1034 • Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix
1035 Hill, Omer Levy, Samuel Bowman. [SuperGLUE: A Stickier Benchmark for General-Purpose](#)
1036 [Language Understanding Systems](#). In *Proceedings of NeurIPS*, 2019. [1/1]
- 1037 • Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska
1038 Roesner, Yejin Choi. [Defending Against Neural Fake News](#). In *Proceedings of NeurIPS*,
1039 2019. [2/4]
- 1040 • Yuan Cao, Quanquan Gu. [Generalization Bounds of Stochastic Gradient Descent for Wide](#)
1041 [and Deep Neural Networks](#). In *Proceedings of NeurIPS*, 2019. [0/1]
- 1042 • Florian Tramer, Dan Boneh. [Adversarial Training and Robustness for Multiple Perturbations](#).
1043 In *Proceedings of NeurIPS*, 2019. [0/2]
- 1044 • Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, Percy S. Liang. [Unla-](#)
1045 [beled Data Improves Adversarial Robustness](#). In *Proceedings of NeurIPS*, 2019. [0/1]
- 1046 • Lars Maaløe, Marco Fraccaro, Valentin Liévin, Ole Winther. [BIVA: A Very Deep Hierarchy](#)
1047 [of Latent Variables for Generative Modeling](#). In *Proceedings of NeurIPS*, 2019. [0/1]
- 1048 • Zeyuan Allen-Zhu, Yuanzhi Li, Yingyu Liang. [Learning and Generalization in Overparam-](#)
1049 [eterized Neural Networks, Going Beyond Two Layers](#). In *Proceedings of NeurIPS*, 2019.
1050 [0/1]
- 1051 • Durk P. Kingma, Prafulla Dhariwal. [Glow: Generative Flow with Invertible 1x1 Convolu-](#)
1052 [tions](#). In *Proceedings of NeurIPS*, 2018. [0/2]
- 1053 • Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, David K. Duvenaud. [Neural Ordinary](#)
1054 [Differential Equations](#). In *Proceedings of NeurIPS*, 2018. [0/1]
- 1055 • Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, Jure Leskovec.
1056 [Hierarchical Graph Representation Learning with Differentiable Pooling](#). In *Proceedings of*
1057 *NeurIPS*, 2018. [0/1]
- 1058 • Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, David K. Duvenaud. [Isolating Sources of](#)
1059 [Disentanglement in Variational Autoencoders](#). In *Proceedings of NeurIPS*, 2018. [0/1]
- 1060 • Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, Baoquan Chen. [PointCNN:](#)
1061 [Convolution On X-Transformed Points](#). In *Proceedings of NeurIPS*, 2018. [0/1]
- 1062 • Arthur Jacot, Franck Gabriel, Clement Hongler. [Neural Tangent Kernel: Convergence and](#)
1063 [Generalization in Neural Networks](#). In *Proceedings of NeurIPS*, 2018. [0/1]
- 1064 • Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan
1065 Catanzaro. [Video-to-Video Synthesis](#). In *Proceedings of NeurIPS*, 2018. [0/1]

- 1066 • Yuanzhi Li, Yingyu Liang. [Learning Overparameterized Neural Networks via Stochastic](#)
1067 [Gradient Descent on Structured Data](#). In *Proceedings of NeurIPS*, 2018. [0/1]
- 1068 • Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, Aleksander Madry.
1069 [Adversarially Robust Generalization Requires More Data](#). In *Proceedings of NeurIPS*, 2018.
1070 [0/2]
- 1071 • Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, Aleksander Madry. [How Does Batch](#)
1072 [Normalization Help Optimization?](#) In *Proceedings of NeurIPS*, 2018. [0/1]
- 1073 • Harini Kannan, Alexey Kurakin, Ian Goodfellow. [Adversarial Logit Pairing](#). In *Proceedings*
1074 *of NeurIPS**, 2018. [0/2]
- 1075 • Ofir Nachum, Shixiang (Shane) Gu, Honglak Lee, Sergey Levine. [Data-Efficient Hierarchi-](#)
1076 [cal Reinforcement Learning](#). In *Proceedings of NeurIPS*, 2018. [0/3]
- 1077 • Prateek Jain, Raghu Meka, Inderjit Dhillon. [Guaranteed Rank Minimization via Singular](#)
1078 [Value Projection](#). In *Proceedings of NeurIPS*, 2010. [0/1]
- 1079 • Hanna Wallach, David Mimno, Andrew McCallum. [Rethinking LDA: Why Priors Matter](#).
1080 In *Proceedings of NeurIPS*, 2009. [0/4]
- 1081 • Geoffrey E. Hinton, Russ R. Salakhutdinov. [Replicated Softmax: an Undirected Topic](#)
1082 [Model](#). In *Proceedings of NeurIPS*, 2009. [0/1]
- 1083 • Daniel J. Hsu, Sham M. Kakade, John Langford, Tong Zhang. [Multi-Label Prediction via](#)
1084 [Compressed Sensing](#). In *Proceedings of NeurIPS*, 2009. [0/1]
- 1085 • Youngmin Cho, Lawrence Saul. [Kernel Methods for Deep Learning](#). In *Proceedings of*
1086 *NeurIPS*, 2009. [0/1]
- 1087 • Kurt Miller, Michael Jordan, Thomas Griffiths. [Nonparametric Latent Feature Models for](#)
1088 [Link Prediction](#). In *Proceedings of NeurIPS*, 2009. [0/3]
- 1089 • Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, Andrew Ng. [Measuring Invariances](#)
1090 [in Deep Networks](#). In *Proceedings of NeurIPS*, 2009. [0/1]
- 1091 • Vinod Nair, Geoffrey E. Hinton. [3D Object Recognition with Deep Belief Nets](#). In *Proceed-*
1092 *ings of NeurIPS*, 2009. [0/1]
- 1093 • Martin Zinkevich, John Langford, Alex Smola. [Slow Learners are Fast](#). In *Proceedings of*
1094 *NeurIPS*, 2009. [0/1]
- 1095 • Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, Gideon Mann. [Efficient](#)
1096 [Large-Scale Distributed Training of Conditional Maximum Entropy Models](#). In *Proceedings*
1097 *of NeurIPS*, 2009. [0/1]
- 1098 • Corinna Cortes, Mehryar Mohri, Afshin Rostamizadeh. [Learning Non-Linear Combinations](#)
1099 [of Kernels](#). In *Proceedings of NeurIPS*, 2009. [0/1]
- 1100 • Laurent Jacob, Jean-philippe Vert, Francis Bach. [Clustered Multi-Task Learning: A Convex](#)
1101 [Formulation](#). In *Proceedings of NeurIPS*, 2008. [0/1]
- 1102 • Kamalika Chaudhuri, Claire Monteleoni. [Privacy-preserving logistic regression](#). In *Proceed-*
1103 *ings of NeurIPS*, 2008. [0/3]
- 1104 • Byron M. Yu, John P. Cunningham, Gopal Santhanam, Stephen Ryu, Krishna V. Shenoy,
1105 Maneesh Sahani. [Gaussian-process factor analysis for low-dimensional single-trial analysis](#)
1106 [of neural population activity](#). In *Proceedings of NeurIPS*, 2008. [0/3]
- 1107 • Ilya Sutskever, Geoffrey E. Hinton, Graham W. Taylor. [The Recurrent Temporal Restricted](#)
1108 [Boltzmann Machine](#). In *Proceedings of NeurIPS*, 2008. [0/1]
- 1109 • Wenyuan Dai, Yuqiang Chen, Gui-rong Xue, Qiang Yang, Yong Yu. [Translated Learning:](#)
1110 [Transfer Learning across Different Feature Spaces](#). In *Proceedings of NeurIPS*, 2008. [0/3]
- 1111 • Yishay Mansour, Mehryar Mohri, Afshin Rostamizadeh. [Domain Adaptation with Multiple](#)
1112 [Sources](#). In *Proceedings of NeurIPS*, 2008. [0/1]
- 1113 • Sham M. Kakade, Karthik Sridharan, Ambuj Tewari. [On the Complexity of Linear Predic-](#)
1114 [tion: Risk Bounds, Margin Bounds, and Regularization](#). In *Proceedings of NeurIPS*, 2008.
1115 [0/1]

- 1116 • Francis Bach. Exploring Large Feature Spaces with Hierarchical Multiple Kernel Learning.
1117 In *Proceedings of NeurIPS*, 2008. [0/1]
- 1118 • Ijaz Akhter, Yaser Sheikh, Sohaib Khan, Takeo Kanade. Nonrigid Structure from Motion in
1119 Trajectory Space. In *Proceedings of NeurIPS*, 2008. [0/1]
- 1120 • Prateek Jain, Brian Kulis, Inderjit Dhillon, Kristen Grauman. Online Metric Learning and
1121 Fast Similarity Search. In *Proceedings of NeurIPS*, 2008. [0/1]
- 1122 • Duy Nguyen-tuong, Jan Peters, Matthias Seeger. Local Gaussian Process Regression for
1123 Real Time Online Model Learning. In *Proceedings of NeurIPS*, 2008. [0/1]
- 1124 • Lester Mackey. Deflation Methods for Sparse PCA. In *Proceedings of NeurIPS*, 2008. [0/1]