Learning Theory of Stochastic Gradient Methods

Yunwen Lei

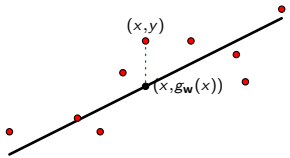
University of Birmingham

AI TIME, September 27, 2021

Background

Supervised Machine Learning

- Given IID training examples from a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
  - ( ,`dog`), ( ,`car`), ( ,`airplane`), ...
  - formally $S = \{z_i = (x_i, y_i), i = 1, \ldots, n\}$, $z_i \in \mathcal{Z}$
- Aim to find prediction rule $g_{\mathbf{w}} : \mathcal{X} \mapsto \mathcal{Y}$, parameterized by $\mathbf{w} \in \Omega$
  - e.g., linear models: $g_{\mathbf{w}}(x) = \langle \mathbf{w}, x \rangle$
- Loss function $f(\mathbf{w}; z)$ to measure performance of $\mathbf{w}$ on an example z
  - least-squares loss: $f(\mathbf{w}; z) = (y - g_{\mathbf{w}}(x))^2$



Population and Empirical Risk

Aim: build a model with small population risk (testing error) $F(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}; z)]$

F is unknown, which is approximated by empirical risk (training error) on S

$$F_S(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{w}; z_i)$$

A learning algorithm A with an output model $A(S) \in \Omega$

- e.g., empirical risk minimization: $A(S) = \arg\min_{\mathbf{w} \in \Omega} F_S(\mathbf{w})$
- regularized risk minimization: $A(S) = \arg\min_{\mathbf{w} \in \Omega} F_S(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2$
- gradient descent (GD)
- stochastic gradient descent (SGD)

Gradient Descent

Gradient Descent (GD)

for $t = 1, 2, \ldots$ **to** T **do**
  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla F_S(\mathbf{w}_t)$ for some step sizes $\eta_t > 0$
return $\mathbf{w}_{T+1}$ or an average of $\mathbf{w}_1, \ldots, \mathbf{w}_{T+1}$

☺ simple, works well for many ML problems

☹ computing $\nabla F_S(\mathbf{w}_t)$ is $O(n)$, slow if n is large

$$\nabla F_S(\mathbf{w}_t) = \frac{1}{n} \sum_{i=1}^{n} \nabla f(\mathbf{w}_t; z_i).$$

GD requires to go through examples for a gradient computation!

Stochastic Gradient Descent

Stochastic Gradient Descent (SGD)

for $t = 1, 2, \ldots$ **to** T **do**
  $i_t \leftarrow$ random index from $\{1, 2, \ldots, n\}$
  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t; z_{i_t})$ for some step sizes $\eta_t > 0$
return $\mathbf{w}_{T+1}$ or an average of $\mathbf{w}_1, \ldots, \mathbf{w}_{T+1}$

- computation cost per iteration is $O(1)$ instead of $O(n)$
- correct in expectation:

$$\mathbb{E}_{i_t}[\nabla f(\mathbf{w}_t; z_{i_t})] = \frac{1}{n} \sum_{i=1}^{n} \nabla f(\mathbf{w}_t; z_i) = \nabla F_S(\mathbf{w}_t)$$

- widely used in training deep neural networks (DNNs)

Theoretical (especially statistical) behavior of SGD is not well understood!

Excess Generalization Error

Let $\mathbf{w}^*$ be the best model parameter

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\Omega} F(\mathbf{w}).$$

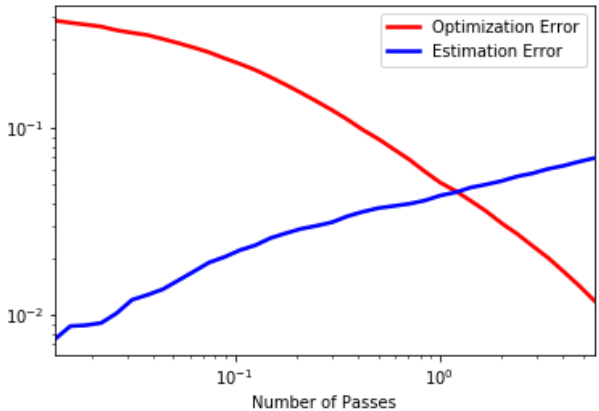
Target of analysis: excess generalization error

$$\mathbb{E}\big[F(A(S)) - F(\mathbf{w}^*)\big] = \mathbb{E}\Big[\underbrace{F(A(S)) - F_S(A(S))}_{\text{estimation error}} + \underbrace{F_S(A(S)) - F_S(\mathbf{w}^*)}_{\text{optimization error}}\Big]$$

1. estimation error: difference between testing error and training error at $A(S)$
2. optimization error: difference between $A(S)$ and $\mathbf{w}^*$ measured by training error

Estimation and Optimization Errors

- Optimization errors decrease as we increase the number of iterations
- Estimation errors increase as we increase the number of iterations
- We need to balance these two errors by early-stopping



Estimation and Optimization Errors

There is a huge literature on optimization errors in optimization theory Bach and Moulines (2013); Smale and Yao (2006); Duchi et al. (2010); Johnson and Zhang (2013); Zhang (2004a); Bottou (1998); Bottou et al. (2018); Shamir and Zhang (2013); Rakhlin et al. (2012); Nemirovski et al. (2009); Nesterov (2015); Ying and Pontil (2008); Ying and Zhou (2017)

There is a huge literature on estimation errors in statistical learning theory Zhou (2002); Shi et al. (2011); Hu et al. (2013); Bartlett et al. (2006); Zhang (2004b); Tsybakov (2004); Vapnik (2013); Bartlett and Mendelson (2002); Lin et al. (2017); Cucker and Zhou (2007); Smale and Zhou (2007); Steinwart and Christmann (2008); Guo et al. (2016)

There is far less study to consider these two errors together Bousquet and Bottou (2008); Hardt et al. (2016); Lin and Rosasco (2017); Yao et al. (2007)

Our contribution: study estimation and optimization error in a framework!

Approaches to Estimation Errors

Stability Approach:

- estimate sensitivity of model wrt perturbation of sample

Hardt et al. (2016); Kuzborskij and Lampert (2018); Charles and Papailiopoulos (2018); Feldman and Vondrak (2019); Bousquet et al. (2020)

Uniform Convergence Approach:

- bound $\sup_{\mathbf{w}\in\Omega} \left|F_S(\mathbf{w}) - F(\mathbf{w})\right|$

Bartlett and Mendelson (2002); Lin et al. (2016)

Integral Operator Approach:

- use the structure of least-square loss

Smale and Zhou (2007); Rosasco and Villa (2015); Ying and Pontil (2008); Lin and Rosasco (2017); Dieuleveut and Bach (2016); Lin and Zhou (2017)

Outline

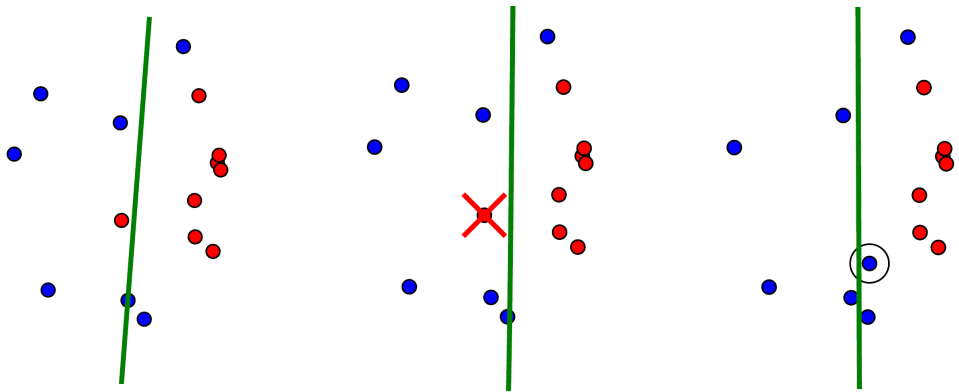
- Stability and Generalization of SGD
- Extensions
- Summary

Stability and Generalization of SGD

Uniform Stability Approach

A randomized algorithm A is ϵ-uniformly stable if, for any two datasets S and S' that differ by one example (neighbor dataset), we have (Bousquet and Elisseeff, 2002; Elisseeff et al., 2005)

$$\sup_z \mathbb{E}_A\big[f(A(S); z) - f(A(S'); z)\big] \leq \epsilon. \tag{1}$$



If A is uniformly stable, then it is generalizable!

Uniform Stability Approach

Existing results

(Hardt et al., 2016)

If f is convex

- strongly smooth, i.e, $\left\|\nabla f(\mathbf{w}, z) - \nabla f(\mathbf{w}', z)\right\|_2 \leq L\|\mathbf{w} - \mathbf{w}'\|_2$
- B-Lipschitz, i.e., $\|\nabla f(\mathbf{w}; z)\|_2 \leq B$

For SGD with step size η_t, informally we have (Let $\{\mathbf{w}_t\}_t$ and $\{\mathbf{w}'_t\}$ be SGD sequences on **neighboring** S and S')

$$\text{estimation error} \leq \text{uniform stability} \leq \underbrace{\mathbb{E}[\|\mathbf{w}_T - \mathbf{w}'_T\|_2]}_{\text{argument stability}} \leq \frac{2B}{n}\sum_{t=1}^{T}\eta_t.$$

The SGD implementation can be represented an **gradient update** defined by

$$G_{\eta,z}(\mathbf{w}) := \mathbf{w} - \eta\nabla f(\mathbf{w}; z) \Longrightarrow \mathbf{w}_{t+1} = G_{\eta,z_{i_t}}(\mathbf{w}_t).$$

Key Property: Let f be convex, L-smooth and B-Lipschitz.

1. If $\eta \leq 2/L$, then $G_{\eta,z}$ is contractive: $\|G_{\eta,z}(\mathbf{w}_t) - G_{\eta,z}(\mathbf{w}'_t)\|_2 \leq \|\mathbf{w}_t - \mathbf{w}'_t\|_2$
2. $G_{\eta,z}$ is B-bounded: $\|G_{\eta,z}(\mathbf{w}) - \mathbf{w}\|_2 \leq B\eta$.

Assumptions are Restrictive

Lipschitz continuity fails for the least square loss

- $f(\mathbf{w}; z) = |\langle \mathbf{w}, x \rangle - y|^2$
- $\nabla f(\mathbf{w}; z) = 2\big(\langle \mathbf{w}, x \rangle - y\big)x$

Smoothness fails for the hinge loss

- $f(\mathbf{w}; z) = \max\big\{0, 1 - y\langle \mathbf{w}, x \rangle\big\}$
- not even differentiable

Can we remove these assumptions and explain the real power of SGD?

Key Idea: Let f be convex, Lipschitz but **not** smooth.

- Standard analysis shows $\|G_{\eta,z}(\mathbf{w}_t) - G_{\eta,z}(\mathbf{w}'_t)\|_2 \le \|\mathbf{w}_t - \mathbf{w}'_t\|_2 + \eta_t$

$$\implies \text{unif stab} \le \sum_{t=1}^{T} \eta_t \implies \text{risk} \le \underbrace{\sum_{t=1}^{T} \eta_t}_{\text{estimation}} + \underbrace{1/(\sum_{t=1}^{T} \eta_t)}_{\text{optimization}}$$

- Question: how about considering $\|G_{\eta,z}(\mathbf{w}_t) - G_{\eta,z}(\mathbf{w}'_t)\|_2^2$ and showing

$$\|G_{\eta,z}(\mathbf{w}_t) - G_{\eta,z}(\mathbf{w}'_t)\|_2^2 \le \|\mathbf{w}_t - \mathbf{w}'_t\|_2^2 + \eta_t^2 \overset{?}{\implies} \text{risk} \le \Big(\sum_{t=1}^{T} \eta_t^2\Big)^{\frac{1}{2}} + 1/(\sum_{t=1}^{T} \eta_t)$$

Let f be convex, smooth but **not** Lipschitz.

- It is clear $\|\mathbf{w} - G_{\eta,z_i}(\mathbf{w})\|_2 = \eta \|\nabla f(\mathbf{w}; z_i)\|_2$
- Self-bounding property of smooth and nonnegative f implies $\|\nabla f(\mathbf{w}; z)\|_2 \le \sqrt{2Lf(\mathbf{w}; z)}$
- Question: how about considering all z_i (boundedness of averaged gradient update)

$$\frac{1}{n\eta(2L)^{\frac{1}{2}}} \sum_{i=1}^{n} \|\mathbf{w} - G_{\eta,z_i}(\mathbf{w})\|_2 \le \Big(\frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}; z_i)\Big)^{\frac{1}{2}} = \big(\text{training error}\big)^{\frac{1}{2}}$$

On-Average Model Stability

To handle the general setting, we propose a new concept of stability.

$$S = \{z_1, z_2, \ldots, z_n\}$$
$$S' = \{z_1', z_2', \ldots, z_n'\}$$

$$\xRightarrow{\text{perturbation}}$$

$$S = \{z_1, z_2, \ldots, z_n\} \xrightarrow{A} A(S)$$
$$S^{(1)} = \{z_1', z_2, \ldots, z_n\} \xrightarrow{A} A(S^{(1)})$$
$$S^{(2)} = \{z_1, z_2', \ldots, z_n\} \xrightarrow{A} A(S^{(2)})$$
$$\vdots$$
$$S^{(n)} = \{z_1, z_2, \ldots, z_n'\} \xrightarrow{A} A(S^{(n)})$$

On-Average Model Stability

We say a randomized algorithm $A : \mathcal{Z}^n \mapsto \Omega$ is on-average model ϵ-stable if

$$\mathbb{E}_{S,S',A}\left[\frac{1}{n}\sum_{i=1}^{n} \|A(S) - A(S^{(i)})\|_2^2\right] \leq \epsilon^2. \tag{2}$$

Y. Lei and Y. Ying. "Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent." *International Conference on Machine Learning*, 2020.

Generalization by On-average Model stability

Hölder Continuous Gradients

We say f has α-Hölder continuous gradients ($\alpha \in [0, 1]$) if

$$\left\|\nabla f(\mathbf{w}, z) - \nabla f(\mathbf{w}', z)\right\|_2 \leq \|\mathbf{w} - \mathbf{w}'\|_2^{\alpha}. \tag{3}$$

- $\alpha = 0$ means that f is Lipschitz and $\alpha = 1$ means strongly smoothness.

Generalization by On-average Model stability

If A is on-average model ϵ-stable, then

$$\text{estimation error} = O\Big(\epsilon^{1+\alpha} + \epsilon\big(\text{training error}\big)^{\frac{\alpha}{1+\alpha}}\Big). \tag{4}$$

- Can handle both Lipschitz functions and un-bounded gradients!
- If training error $= 0$, then estimation error $= O\big(\epsilon^{1+\alpha}\big)$.
- This is much *faster* than estimation error $= O(\epsilon)$.

Main Results for SGD

On-Average Model Stability for SGD

- If ∇f is α-Hölder continuous with $\alpha \in [0, 1]$, then

$$\epsilon_{T+1}^2 = O\Big(\sum_{t=1}^{T} \eta_t^{\frac{2}{1-\alpha}} + \frac{1 + T/n}{n} \big(\sum_{t=1}^{T} \eta_t^2 \big)^{\frac{1-\alpha}{1+\alpha}} \Big(\sum_{t=1}^{T} \eta_t^2 \mathbb{E}[F_S(\mathbf{w}_t)] \Big)^{\frac{2\alpha}{1+\alpha}} \Big)$$

- *Weighted sum of training errors* (i.e. $\sum_{t=1}^{T} \eta_t^2 \mathbb{E}\big[F_S(\mathbf{w}_t)\big]$) can be estimated using tools of analyzing optimization errors

Estimation error $\leq$ On-average model stability $\leq$ Weighted sum of training errors

Recall, for uniform stability with Lipschitz and smooth f, that

$$\text{Estimation error} \leq \text{Uniform stability} \leq \frac{2B}{n} \sum_{t=1}^{T} \eta_t$$

SGD with Smooth Functions

Let f be convex and strongly-smooth. Let $\bar{\mathbf{w}}_T = \sum_{t=1}^{T} \eta_t \mathbf{w}_t / \sum_{t=1}^{T} \eta_t$.

Theorem (Minimax optimal generalization bounds)

Choosing $\eta_t = 1/\sqrt{T}$ and $T \asymp n$ implies that

$$\mathbb{E}\big[F(\bar{\mathbf{w}}_T)\big] - F(\mathbf{w}^*) = O\big(1/\sqrt{n}\big).$$

Theorem (Fast generalization bounds under low noise)

For low noise case $F(\mathbf{w}^*) = O(1/n)$, we can take $\eta_t = 1$, $T \asymp n$ and get

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] = O(1/n).$$

- We remove bounded gradient assumptions.
- We get the first-ever fast generalization bound $O(1/n)$ by stability analysis.

SGD with Lipschitz Functions

Let f be convex and G-Lipschitz (Not necessarily smooth! e.g. the hinge loss.)

Our on-average model stability bounds can be simplified as

$$\epsilon_{T+1}^2 = O\Big((1 + T/n^2) \sum_{t=1}^{T} \eta_t^2\Big). \tag{5}$$

Key idea: gradient update is approximately contractive

$$\|G_{\eta,z}(\mathbf{w}) - G_{\eta,z}(\mathbf{v})\|_2^2 \leq \|\mathbf{w} - \mathbf{v}\|_2^2 + O(\eta^2). \tag{6}$$

Theorem (Generalization bounds)

We can take $\eta_t = T^{-\frac{3}{4}}$ *and* $T \asymp n^2$ *and get*

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}}).$$

We get the first generalization bound $O(1/\sqrt{n})$ for SGD with non-differentiable functions based on stability analysis.

SGD with α-Hölder continuous gradients

Let f be convex and have α-Hölder continuous gradients with $\alpha \in (0, 1)$.

Key idea: gradient update is approximately contractive

$$\|G_{\eta,z}(\mathbf{w}) - G_{\eta,z}(\mathbf{v})\|_2^2 \leq \|\mathbf{w} - \mathbf{w}'\|_2^2 + O(\eta^{\frac{2}{1-\alpha}}).$$

Theorem

- *If* $\alpha \geq 1/2$, *we take* $\eta_t = 1/\sqrt{T}$, $T \asymp n$ *and get*

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}}).$$

- *If* $\alpha < 1/2$, *we take* $\eta_t = T^{\frac{3\alpha-3}{2(2-\alpha)}}$, $T \asymp n^{\frac{2-\alpha}{1+\alpha}}$ *and get*

$$\mathbb{E}[F(\bar{\mathbf{w}}_T)] - F(\mathbf{w}^*) = O(n^{-\frac{1}{2}}).$$

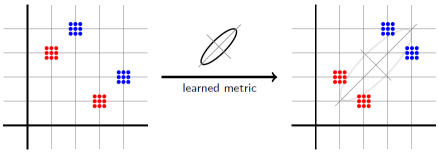
Theorem (Fast Generalization bounds)

If $F(\mathbf{w}^*) = O(\frac{1}{n})$, *we let* $\eta_t = T^{\frac{\alpha^2+2\alpha-3}{4}}$, $T \asymp n^{\frac{2}{1+\alpha}}$ *and get* $\mathbb{E}[F(\bar{\mathbf{w}}_T)] = O(n^{-\frac{1+\alpha}{2}})$.

Extensions

Stochastic Gradient Methods for Pairwise Learning

- Pairwise learning unifies several important problems



(a): Metric Learning

(b): Ranking

- Pairwise loss: $f(\mathbf{w}; z, z')$ measures behavior of $h_{\mathbf{w}}$ over z, z'

$$\text{testing error} \quad F(\mathbf{w}) = \mathbb{E}_{z,z'}\big[f(\mathbf{w}; z, z')\big], \quad \text{training error} \quad F_S(\mathbf{w}) = \frac{1}{n(n-1)} \sum_{i,j\in[n]:i\neq j} f(\mathbf{w}; z_i, z_j).$$

- Note the summands in $F_S(\mathbf{w})$ are not independent, e.g., $f(\mathbf{w}; z_1, z_2)$ and $f(\mathbf{w}; z_1, z_3)$

We propose novel pairwise learning algorithms and develop a framework to study stability, generalization and optimization!

Y. Lei, A. Ledent and M. Kloft. "Sharper Generalization Bounds for Pairwise Learning." *Advances in Neural Information Processing Systems*, pages 21236-21246, 2020.

Stability and Generalization for Minimax Problems

- Minimax formulation (e.g. GAN, AUC maximization, and robust learning):

$$\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{v}\in\mathcal{V}}\Big\{F(\mathbf{w},\mathbf{v}) := \mathbb{E}_z[f(\mathbf{w},\mathbf{v};z)]\Big\}. \tag{7}$$

- In practice, a randomized optimization algorithm A (e.g. SGDA) is employed to solve its empirical version:

$$\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{v}\in\mathcal{V}}\Big\{F_S(\mathbf{w},\mathbf{v}) := \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w},\mathbf{v};z_i)\Big\}. \tag{8}$$

- The literature is vast and most of them focused on the convergence of the output of A, i.e.

$$A(S) = (A_{\mathbf{w}}(S), A_{\mathbf{v}}(S)).$$

We develop a framework to study the stability and generalization for stochastic gradient methods for minimax problems!

Y. Lei, Z. Yang, T. Yang and Y. Ying "Stability and Generalization of Stochastic Gradient Methods for Minimax Problems." In *International Conference on Machine Learning*, pages 6175-6186, 2021.

Conclusion

Summary

Stability and Generalization of SGD

- novel stability measures
- remove restrictive assumptions
- better bounds

Extensions

- pairwise learning
- minimax problems

References I

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

L. Bottou. On-line learning and stochastic approximations. In D. Saad, editor, *On-line Learning in Neural Networks*, pages 9–42. Cambridge University Press, New York, NY, USA, 1998.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. pages 610–626, 2020.

Z. Charles and D. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International Conference on Machine Learning*, pages 744–753, 2018.

F. Cucker and D.-X. Zhou. *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press, 2007.

A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Conference on Learning Theory*, page 257, 2010.

A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(Jan):55–79, 2005.

V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279, 2019.

X. Guo, J. Fan, and D.-X. Zhou. Sparsity and error analysis of empirical feature-based regularization schemes. *Journal of Machine Learning Research*, 17(89):1–34, 2016.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

T. Hu, J. Fan, Q. Wu, and D.-X. Zhou. Learning theory approach to minimum error entropy criterion. *Journal of Machine Learning Research*, 14(Feb):377–397, 2013.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

I. Kuzborskij and C. Lampert. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 2820–2829, 2018.

Y. Lei and Y. Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819, 2020.

Y. Lei, A. Ledent, and M. Kloft. Sharper generalization bounds for pairwise learning. *Advances in Neural Information Processing Systems*, 33:21236–21246, 2020.

Y. Lei, Z. Yang, T. Yang, and Y. Ying. Stability and generalization of stochastic gradient methods for minimax problems. In *International Conference on Machine Learning*, pages 6175–6186, 2021.

J. Lin and L. Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(1):3375–3421, 2017.

J. Lin, R. Camoriano, and L. Rosasco. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning*, pages 2340–2348, 2016.

S.-B. Lin and D.-X. Zhou. Distributed kernel-based gradient descent algorithms. *Constructive Approximation*, pages 1–28, 2017.

References II

S.-B. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. *The Journal of Machine Learning Research*, 18(1):3202–3232, 2017.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Y. Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pages 449–456, 2012.

L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

L. Shi, Y.-L. Feng, and D.-X. Zhou. Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31(2):286–302, 2011.

S. Smale and Y. Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6(2):145–170, 2006.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

Y. Ying and D.-X. Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224—-244, 2017.

T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning*, pages 919–926, 2004a.

T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004b.

D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Thank you! AI TIME