Stochastic Gradient Descent: Algorithmic Stability and Implicit Regularization[1]

Yunwen Lei

Hong Kong Baptist University

6th Young Scholar Symposium
East Asia Section of Inverse Problems International Association
25 March, 2023

[1] Joint work with Yiming Ying

Background

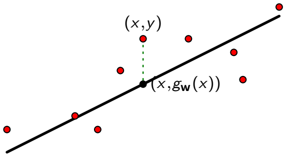
Supervised Machine Learning

- Given training examples from a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
  - (,dog), (,car), (,airplane), ...
  - formally $S = \{z_i = (x_i, y_i), i = 1, \ldots, n\}$, $z_i \in \mathcal{Z}$
  - Independently drawn from a probability measure ρ on $\mathcal{Z}$
- Aim to find prediction rule $g_{\mathbf{w}} : \mathcal{X} \mapsto \mathcal{Y}$, parameterized by $\mathbf{w} \in \mathcal{W}$ (model space)
  - linear models: $g_{\mathbf{w}}(x) = \langle \mathbf{w}, x \rangle$
  - neural networks: $g_{\mathbf{w}}(x) = \sigma_L(\mathbf{W}_L \sigma_{L-1}(\mathbf{W}_{L-1} \cdots \sigma_1(\mathbf{W}_1 x)))$

Population and Empirical Risk

Loss function $f(\mathbf{w}; z)$ to measure performance of $g_{\mathbf{w}}$ on an example $z = (x, y)$

- squares loss: $f(\mathbf{w}; z) = (y - g_{\mathbf{w}}(x))^2$ for regression



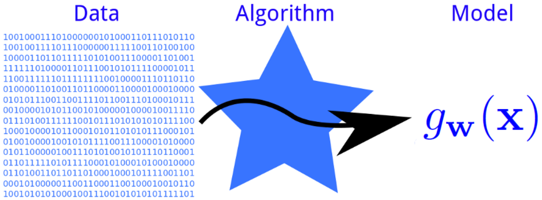
- hinge loss: $f(\mathbf{w}; z) = \max\{0, 1 - y g_{\mathbf{w}}(x)\}$ for binary classification

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

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

Algorithms



- A learning algorithm A with an output model $A(S) \in \mathcal{W}$
  - empirical risk minimization: $A(S) = \arg\min_{\mathbf{w} \in \mathcal{W}} \text{training\_error}(\mathbf{w})$
  - regularized risk minimization:

$$A(S) = \arg\min_{\mathbf{w} \in \mathcal{W}} \{\text{training\_error}(\mathbf{w}) + \text{regularizer}(\mathbf{w})\}$$

  - gradient descent, stochastic gradient descent, stochastic gradient descent ascent ...

Generalization Gap

- Algorithm A often produces models with a small training error
- This does not necessarily mean $A(S)$ has a good prediction
- This asks for the study of an important concept called **generalization gap**

$$\textbf{Generalization gap} = \text{Test Error} - \text{Training Error}$$

Our work: Statistics + Optimization

We focus on generalization issues of optimization algorithms via algorithmic stability

- implicit regularization (no regularizer in the objective function)
- how to trade off optimization and generalization for good prediction

Stability and Generalization of 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 Population Risk

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

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

Target of analysis: excess population risk

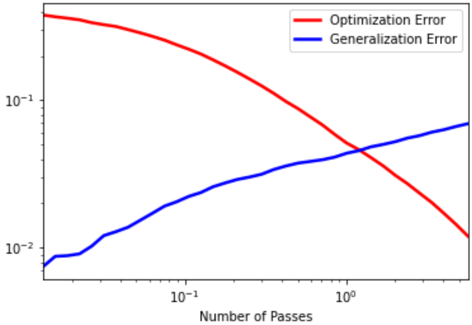
$$\mathbb{E}\big[F(A(S)) - F(\mathbf{w}^*)\big] = \mathbb{E}\Big[\underbrace{F(A(S)) - F_S(A(S))}_{\text{generalization gap}} + \underbrace{F_S(A(S)) - F_S(\mathbf{w}^*)}_{\text{optimization error}}\Big]$$

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

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

Generalization and Optimization Errors

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



Generalization and Optimization Errors

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

There is a huge literature on generalization gap in statistical learning theory

- **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 $\sup_{\mathbf{w}\in\mathcal{W}} \big|F_S(\mathbf{w}) - F(\mathbf{w})\big|$ (Zhang, 2004b; Zhou, 2002; Cucker and Smale, 2002; Bartlett and Mendelson, 2002; Lin et al., 2016; Tsybakov, 2004; Cucker and Zhou, 2007; Vapnik, 2013; Steinwart and Christmann, 2008)
- **Integral Operator** Approach: use the structure of square loss (Smale and Zhou, 2007; Rosasco and Villa, 2015; Ying and Pontil, 2008; Lin and Rosasco, 2017; Dieuleveut and Bach, 2016; Lin et al., 2017; Lin and Zhou, 2017; Jin et al., 2021)

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 work: study generalization and optimization error in a framework!

Uniform Stability Approach

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

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

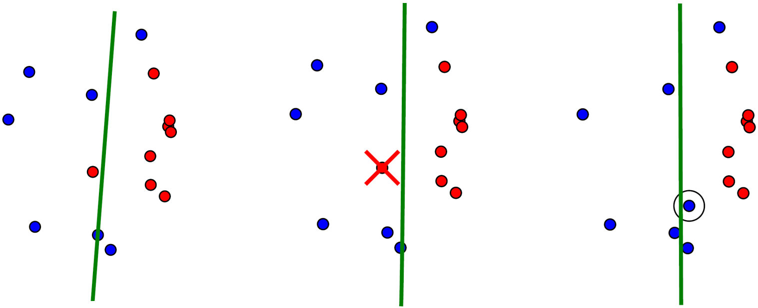


Figure Taken in Kuzborskij and Lampert (2018)

If A is uniformly stable, then it is generalizable!

- if $z \in S' \backslash S$, then z is a test point for $A(S)$ and a training point for $A(S')$
- $f(A(S);z)$ is testing error and $f(A(S');z)$ is training error

Uniform Stability Approach

Existing results

(Hardt et al., 2016)

Let $\{\mathbf{w}_t\}_t$ and $\{\mathbf{w}'_t\}$ be SGD sequences on **neighboring** S and S'. Let f be convex

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

For SGD with step size η_t, informally we have

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

Assumptions are Restrictive

Lipschitz continuity fails for the 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

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

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

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 $A : \mathcal{Z}^n \mapsto \mathcal{W}$ is on-average model ϵ-stable if

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

Generalization by On-average Model stability

Hölder Continuous Gradients

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

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

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

Generalization by On-average Model stability

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

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

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

Main Results for SGD

On-Average Model Stability for SGD

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

$$\epsilon_{T+1}^2 = O\Big(\sum_{t=1}^{T} \eta_t^{\frac{2}{1-\alpha}} + \frac{1+T/n}{n}\big(\sum_{t=1}^{T} \eta_t^2\big)^{\frac{1-\alpha}{1+\alpha}} \Big(\sum_{t=1}^{T} \eta_t^2 \mathbb{E}[F_S(\mathbf{w}_t)]\Big)^{\frac{2\alpha}{1+\alpha}}\Big) \tag{5}$$

- *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

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

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

$$\text{generalization gap} \leq \text{uniform stability} \leq \frac{2B}{n}\sum_{t=1}^{T} \eta_t \tag{6}$$

SGD with Smooth and Convex Functions

Stability bound: $\epsilon_T^2 = O\Big(\frac{1}{n}\sum_{t=1}^T \eta_t^2 \mathbb{E}[F_S(\mathbf{w}_t)]\Big) \Longrightarrow$ generalization bound

Implicit Regularization

Let $A(S)$ be the model given by SGD with $\eta_t = \eta$. There is $C > 0$ such that

$$\mathbb{E}\big[F(A(S))\big] = \min_{\mathbf{w}} \Big\{F(\mathbf{w}) + \frac{C\|\mathbf{w}\|_2^2}{\eta T} + C\eta F(\mathbf{w})\Big\}.$$

SGD actually finds a minimizer of the L_2-regularization with $\lambda = \frac{1}{\eta T}$!

- Choosing $\eta_t = 1/\sqrt{T}$ and $T \asymp n$ implies $\mathbb{E}\big[F(\bar{\mathbf{w}}_T)\big] - F(\mathbf{w}^*) = O\big(1/\sqrt{n}\big)$
- Under a low noise condition $F(\mathbf{w}^*) = 0$, we can take $\eta_t = 1$, $T \asymp n$ and get the first-ever fast bound $O(1/n)$ by stability analysis: $\mathbb{E}[F(A(S))] = O(1/n)$.
- We remove bounded gradient assumptions.

SGD with Lipschitz and Convex Functions

On-average model stability bounds are simplified as $\epsilon_{T+1}^2 = O\Big(\big(1+T/n^2\big)\sum_{t=1}^T \eta_t^2\Big)$.

Key idea: gradient update is approximately nonexpansive

$$\big\|\big(\mathbf{w}-\eta\nabla f(\mathbf{w};z)\big)-\big(\mathbf{w}'-\eta\nabla f(\mathbf{w}';z)\big)\big\|_2^2 = \|\mathbf{w}-\mathbf{w}'\|_2^2 + O(\eta^2). \tag{7}$$

Implicit Regularization

Let $A(S)$ be the model given by SGD with $\eta_t = \eta$. There are C_1, C_2 such that

$$\mathbb{E}\big[F(A(S))\big] = \min_{\mathbf{w}}\Big\{F(\mathbf{w}) + C_1(T\eta)^{-1}\|\mathbf{w}\|_2^2\Big\} + C_2\eta\big(\sqrt{T}+T/n\big).$$

SGD actually finds a minimizer of the L_2-regularization with $\lambda = \frac{1}{T\eta}$!

- 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 risk 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 nonexpansive

$$\big\|\big(\mathbf{w}-\eta\nabla f(\mathbf{w};z)\big)-\big(\mathbf{w}'-\eta\nabla f(\mathbf{w}';z)\big)\big\|_2^2=\|\mathbf{w}-\mathbf{w}'\|_2^2+O(\eta^{\frac{2}{1-\alpha}}).$$

Theorem (Excess risk bounds)

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

Extension

Complexity Analysis of SGD in a Convex Setting

Complexity bound: If $\sum_{t=1}^{\infty} \eta_t^2 < \infty$, then with high probability

$$\max_{t=1,\ldots,T} \|\mathbf{w}_t\|_2 = \widetilde{O}\Big(\frac{1}{\sqrt{n}} \sum_{t=1}^{T} \eta_t\Big).$$

Generalization bound: If $\sum_{t=1}^{\infty} \eta_t^2 < \infty$, then with high probability

$$\max_{t=1,\ldots,T} \big[F(\mathbf{w}_t) - F_S(\mathbf{w}_t)\big] = \widetilde{O}\Big(\frac{1}{n} \sum_{t=1}^{T} \eta_t\Big).$$

Excess risk bound: If $T \asymp n$ and $\eta_t = \widetilde{O}(1/\sqrt{t})$, then with high probability

$$F(\mathbf{w}_T) - F(\mathbf{w}^*) = \widetilde{O}(1/\sqrt{n}).$$

- High probability risk bound for SGD!
- Implicit regularization is achieved by tuning the number of passes and the step size
- No bounded gradient & smoothness assumptions and extended to kernel methods
- Fast rates can be obtained under capacity assumption

Y. Lei, T. Hu and K. Tang. "Generalization Performance of Multi-pass Stochastic Gradient Descent with Convex Loss Functions." *Journal of Machine Learning Research*, 22(25):1-41, 2021.

Stability and Generalization for Non-convex Learning

We assume training errors are gradient-dominated (can be non-convex)

$$\mathbb{E}\big[F_S(\mathbf{w}) - \min_{\mathbf{w}} F_S(\mathbf{w})\big] \leq \frac{1}{2\beta}\mathbb{E}\big[\|\nabla F_S(\mathbf{w})\|_2^2\big], \quad \forall \mathbf{w} \in \mathcal{W}. \tag{8}$$

Examples of gradient-dominated functions are found in dictionary learning, matrix completion, neural networks, etc (Arora et al., 2015; Sun and Luo, 2016; Allen-Zhu et al., 2019)

Theorem (Generalization bounds)

If F_S satisfies (8) and f is smooth, then

$$\text{generalization gap} \leq \text{stability} \leq \frac{1}{n\beta} + \frac{\text{optimization error}}{\beta}. \tag{9}$$

- It applies to **any** algorithm: SGD, SVRG, ADAM...
- Optimization helps generalization: run A until optimization error $\leq 1/n$
- Regularizer is not required for gradient-dominate problems

Y. Lei and Y. Ying. "Sharper Generalization Bounds for Learning with Gradient-dominated Objective Functions." *In International Conference on Learning Representations*, 2021.

Conclusion

Summary

Stability analysis of SGD

- novel stability measures
- remove restrictive assumptions
- better generalization bounds
- implicit regularization

Extensions

- complexity approach
- non-convex learning

References I

Z. Allen-Zhu, Y. Li, and Z. Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. In *Conference on learning theory*, pages 113–149. PMLR, 2015.

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.

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. In *Conference on Learning Theory*, 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 S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

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.

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.

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.

B. Jin, Z. Zhou, and J. Zou. On the saturation phenomenon of stochastic gradient descent for linear inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 9(4):1553–1588, 2021.

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.

References II

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 and Y. Ying. Sharper generalization bounds for learning with gradient-dominated objective functions. In *International Conference on Learning Representations*, 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.

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.

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.

R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

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!