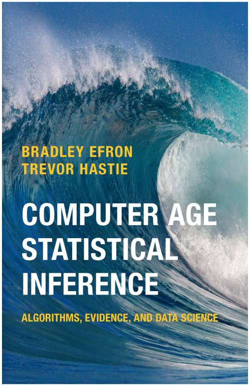
The Implicit Regularization of Stochastic Gradient Flow for Least Squares

Alnur Ali[1], Edgar Dobriban[2], and Ryan J. Tibshirani[3]

[1]Stanford University, [2]University of Pennsylvania,
[3]Carnegie Mellon University

Outline

Introduction



1

Algorithms and Inference

Statistics is the science of learning from experience, particularly experience that arrives a little bit at a time: the successes and failures of a new experimental drug, the uncertain measurements of an asteroid's path toward Earth. It may seem surprising that any one theory can cover such an amorphous target as "learning from experience." In fact, there are *two* main statistical theories, Bayesianism and frequentism, whose connections and disagreements animate many of the succeeding chapters.

Introduction

Given the sizes of modern data sets, stochastic gradient descent is one of the most widely used optimization algorithms today

- Computational and statistical properties have been studied for decades (Robbins & Monro, 1951; Fabian, 1968; Ruppert, 1988; Kushner & Yin, 2003; Polyak & Juditsky, 1992; ...)

Introduction

Given the sizes of modern data sets, stochastic gradient descent is one of the most widely used optimization algorithms today

- Computational and statistical properties have been studied for decades (Robbins & Monro, 1951; Fabian, 1968; Ruppert, 1988; Kushner & Yin, 2003; Polyak & Juditsky, 1992; ...)

Recent interest in implicit regularization [Mahoney and Orecchia, 2011, Mahoney, 2012, Gleich and Mahoney, 2014, Martin and Mahoney, 2018, Soudry et al., 2018, Rosasco and Villa, 2015, Lin et al., 2016, Lin and Rosasco, 2017, Neu and Rosasco, 2018] etc

Introduction

Given the sizes of modern data sets, stochastic gradient descent is one of the most widely used optimization algorithms today

- Computational and statistical properties have been studied for decades (Robbins & Monro, 1951; Fabian, 1968; Ruppert, 1988; Kushner & Yin, 2003; Polyak & Juditsky, 1992; ...)

Recent interest in implicit regularization [Mahoney and Orecchia, 2011, Mahoney, 2012, Gleich and Mahoney, 2014, Martin and Mahoney, 2018, Soudry et al., 2018, Rosasco and Villa, 2015, Lin et al., 2016, Lin and Rosasco, 2017, Neu and Rosasco, 2018] etc

In particular, a line of work showing (early-stopped) gradient descent is linked to ℓ_2 regularization [Nacson et al., 2018, Gunasekar et al., 2018, Suggala et al., 2018, Ali et al., 2018, Poggio et al., 2019, Ji and Telgarsky, 2019] etc

Introduction

Given the sizes of modern data sets, stochastic gradient descent is one of the most widely used optimization algorithms today

- Computational and statistical properties have been studied for decades (Robbins & Monro, 1951; Fabian, 1968; Ruppert, 1988; Kushner & Yin, 2003; Polyak & Juditsky, 1992; ...)

Recent interest in implicit regularization [Mahoney and Orecchia, 2011, Mahoney, 2012, Gleich and Mahoney, 2014, Martin and Mahoney, 2018, Soudry et al., 2018, Rosasco and Villa, 2015, Lin et al., 2016, Lin and Rosasco, 2017, Neu and Rosasco, 2018] etc

In particular, a line of work showing (early-stopped) gradient descent is linked to ℓ_2 regularization [Nacson et al., 2018, Gunasekar et al., 2018, Suggala et al., 2018, Ali et al., 2018, Poggio et al., 2019, Ji and Telgarsky, 2019] etc

"Double win"

Introduction

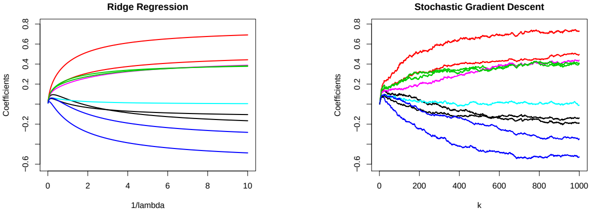
Natural to ask: do the iterates generated by (mini-batch) stochastic gradient descent also possess (implicit) ℓ_2 regularity?

Introduction

Natural to ask: do the iterates generated by (mini-batch) stochastic gradient descent also possess (implicit) ℓ_2 regularity?

Why might there be a connection, at all?

Compare the paths for least squares regression



Here we focus on least squares regression

Introduction

Main tool for making the connection: a stochastic differential equation that we call stochastic gradient flow

Linked to SGD with a constant step size; more on this later

We give a bound on the excess risk of stochastic gradient flow at time t, over ridge regression with tuning parameter $\lambda = 1/t$

Result(s) hold across the entire optimization path
Results do not place strong conditions on the features
Proofs are simpler than in discrete-time (hard to handle var.)

Outline

Stochastic gradient flow

Stochastic gradient flow

Least squares regression (response $y \in \mathbb{R}^n$; data $X \in \mathbb{R}^{n \times p}$)

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2n}\|y - X\beta\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 := F(\beta)$$

Stochastic gradient flow

Least squares regression (response $y \in \mathbb{R}^n$; data $X \in \mathbb{R}^{n \times p}$)

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2n}\|y - X\beta\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 := F(\beta)$$

Mini-batch SGD: $k = 1, 2, 3, \ldots$, $\eta > 0$ fixed step size, m mini-batch size w/ replacement, $\beta^{(0)} = 0$

$$\beta^{(k)} = \beta^{(k-1)} + \frac{\eta}{m} \cdot \sum_{i \in \mathcal{I}_k} (y_i - x_i^\top \beta^{(k-1)}) x_i$$

Stochastic gradient flow

Least squares regression (response $y \in \mathbb{R}^n$; data $X \in \mathbb{R}^{n \times p}$)

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \;\; \frac{1}{2n}\|y - X\beta\|_2^2 = \frac{1}{n}\sum_{i=1}^n (y_i - x_i^\top \beta)^2 := F(\beta)$$

Mini-batch SGD: $k = 1, 2, 3, \ldots$, $\eta > 0$ fixed step size, m mini-batch size w/ replacement, $\beta^{(0)} = 0$

$$\beta^{(k)} = \beta^{(k-1)} + \frac{\eta}{m} \cdot \sum_{i \in \mathcal{I}_k} (y_i - x_i^\top \beta^{(k-1)}) x_i$$

Add+subtract gradient:

$$\begin{aligned} \beta^{(k)} &= \beta^{(k-1)} + \frac{\eta}{n} \cdot X^\top (y - X\beta^{(k-1)}) \\ &+ \eta \cdot \left(\frac{1}{m} X_{\mathcal{I}_k}^\top (y_{\mathcal{I}_k} - X_{\mathcal{I}_k} \beta^{(k-1)}) - \frac{1}{n} X^\top (y - X\beta^{(k-1)}) \right). \end{aligned}$$

Stochastic gradient flow

Motivates the stochastic differential equation [Mandt et al., 2015, Hu et al., 2017, Feng et al., 2017, Li et al., 2019, Feng et al., 2019]

$$d\beta_t = \underbrace{\frac{1}{n} X^\top (y - X\beta_t)\, dt}_{\substack{\text{just the gradient for} \\ \text{least squares regression}}} + \underbrace{Q_\eta(\beta_t)^{1/2}\, dW(t)}_{\substack{\text{fluctuations are governed by the} \\ \text{cov. of the stochastic gradients}}}, \tag{1}$$

where $\beta_0 = 0$,

$$Q_\eta(\beta) = \eta \cdot \mathrm{Cov}_{\mathcal{I}} \left(\frac{1}{m} X_{\mathcal{I}}^\top (y_{\mathcal{I}} - X_{\mathcal{I}} \beta) \right)$$

is the diffusion coefficient, $\mathcal{I} \subseteq \{1, \ldots, n\}$ is a mini-batch, and $\eta > 0$ is a (fixed) step size

We call (1) stochastic gradient flow

Has a few nice properties, and bears several connections to SGD with a constant step size; more on this next

Stochastic gradient flow

Lemma: the Euler discretization of stochastic gradient flow $\tilde{\beta}^{(k)}$, and constant step size SGD $\beta^{(k)}$, share first and second moments, i.e.,

$$\mathbb{E}(\tilde{\beta}^{(k)}) = \mathbb{E}(\beta^{(k)}) \quad \text{and} \quad \mathrm{Cov}(\tilde{\beta}^{(k)}) = \mathrm{Cov}(\beta^{(k)})$$

Stochastic gradient flow

Lemma: the Euler discretization of stochastic gradient flow $\tilde{\beta}^{(k)}$, and constant step size SGD $\beta^{(k)}$, share first and second moments, i.e.,

$$\mathbb{E}(\tilde{\beta}^{(k)}) = \mathbb{E}(\beta^{(k)}) \quad \text{and} \quad \mathrm{Cov}(\tilde{\beta}^{(k)}) = \mathrm{Cov}(\beta^{(k)})$$

Implies the estimation & prediction errors match

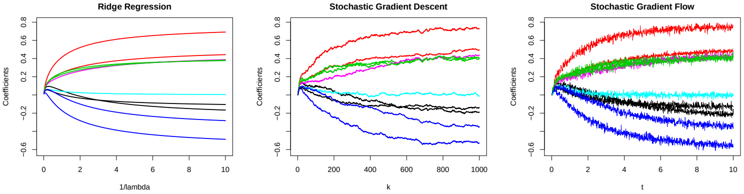
Stochastic gradient flow

Lemma: the Euler discretization of stochastic gradient flow $\tilde{\beta}^{(k)}$, and constant step size SGD $\beta^{(k)}$, share first and second moments, i.e.,

$$\mathbb{E}(\tilde{\beta}^{(k)}) = \mathbb{E}(\beta^{(k)}) \quad \text{and} \quad \mathrm{Cov}(\tilde{\beta}^{(k)}) = \mathrm{Cov}(\beta^{(k)})$$

Implies the estimation & prediction errors match

Sanity check: revisiting the solution/optimization paths from earlier



Stochastic gradient flow

A number of works consider instead the constant covariance process,

$$d\beta_t = \frac{1}{n} X^\top (y - X\beta_t)\, dt + \boxed{\left(\frac{\eta}{m} \cdot \hat{\Sigma}\right)^{1/2} dW(t)}, \qquad (2)$$

where $\hat{\Sigma} = X^\top X/n$ [Mandt et al., 2017, Wang, 2017, Dieuleveut et al., 2017, Fan et al., 2018]

Stochastic gradient flow

A number of works consider instead the constant covariance process,

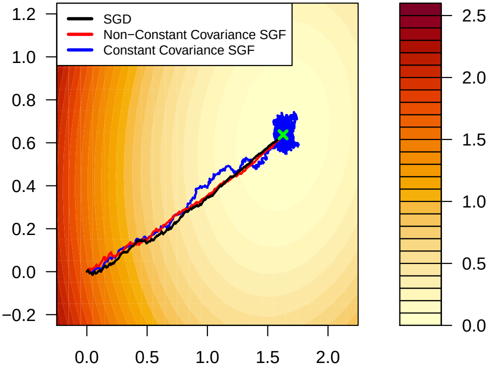
$$d\beta_t = \frac{1}{n} X^\top (y - X\beta_t)\, dt + \left(\frac{\eta}{m} \cdot \hat{\Sigma}\right)^{1/2} dW(t), \qquad (2)$$

where $\hat{\Sigma} = X^\top X/n$ [Mandt et al., 2017, Wang, 2017, Dieuleveut et al., 2017, Fan et al., 2018]

Stochastic Gradient Langevin dynamics (SGLD) takes diffusion coeff as τI [Geman and Hwang, 1986, Seung et al., 1992, Neal, 2011, Welling and Teh, 2011, Sato and Nakagawa, 2014, Teh et al., 2016, Raginsky et al., 2017, Cheng et al., 2019]

Stochastic gradient flow

Stochastic gradient flow is a more accurate approximation



Outline

Statistical Setup

Assume a standard regression model

$$y = X\beta_* + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I)$$

Fix X; let $s_i, i = 1, \ldots, p$, denote the eigenvalues of $X^\top X/n$

Statistical Setup

Assume a standard regression model

$$y = X\beta_* + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I)$$

Fix X; let $s_i, i = 1, \ldots, p$, denote the eigenvalues of $X^\top X/n$

Recall a useful result for (batch) gradient flow (Ali et al., 2018)

For least squares regression, gradient flow is

$$\dot{\beta}_t = \frac{1}{n} X^\top (y - X\beta_t) dt, \quad \beta_0 = 0$$

Has the solution

$$\hat{\beta}_t^{\mathrm{gf}} = (X^\top X)^+ (I - \exp(-tX^\top X/n)) X^\top y$$

Statistical Setup

Assume a standard regression model

$$y = X\beta_* + \varepsilon, \quad \varepsilon \sim (0, \sigma^2 I)$$

Fix X; let $s_i, i = 1, \ldots, p$, denote the eigenvalues of $X^\top X/n$

Recall a useful result for (batch) gradient flow (Ali et al., 2018)

For least squares regression, gradient flow is

$$\dot{\beta}_t = \frac{1}{n} X^\top (y - X\beta_t) dt, \quad \beta_0 = 0$$

Has the solution

$$\hat{\beta}_t^{\mathrm{gf}} = (X^\top X)^+ (I - \exp(-tX^\top X/n)) X^\top y$$

Then, for any time $t \geq 0$ (note the correspondence with λ),

$\mathrm{Bias}^2(\hat{\beta}_t^{\mathrm{gf}}; \beta_*) \leq \mathrm{Bias}^2(\hat{\beta}_{1/t}^{\mathrm{ridge}}; \beta_*)$ and

$\mathrm{Var}(\hat{\beta}_t^{\mathrm{gf}}) \leq 1.6862 \cdot \mathrm{Var}(\hat{\beta}_{1/t}^{\mathrm{ridge}})$, so that

$\mathrm{Risk}(\hat{\beta}_t^{\mathrm{gf}}; \beta_*) \leq 1.6862 \cdot \mathrm{Risk}(\hat{\beta}_{1/t}^{\mathrm{ridge}}; \beta_*)$

Excess risk bound (over ridge)

Thm.: for any time $t > 0$ (provided the step size is small enough),

$$
\begin{aligned}
&\mathrm{Risk}(\hat\beta_t^{\mathrm{sgf}};\beta_*) - \mathrm{Risk}(\hat\beta_{1/t}^{\mathrm{ridge}};\beta_*) \\
&\quad \le 0.6862 \cdot \mathrm{Var}_\varepsilon(\hat\beta_{1/t}^{\mathrm{ridge}}) && \text{(scaled ridge variance)} \\
&\qquad + \eta \cdot \frac{n}{m} \sum_{i=1}^{p} \mathbb{E}_\varepsilon\left[\frac{\exp(\delta_y) s_i}{s_i - \alpha/2}\big(\exp(-\alpha t) - \exp(-2ts_i)\big)\right] \\
&&& \text{(``price of stochasticity'')} \\
&\qquad + \eta \cdot \frac{n}{m} \sum_{i=1}^{p} \mathbb{E}_\varepsilon\Big[\gamma_y\big(1 - \exp(-2ts_i)\big)\Big] && \text{(limiting opt. error)}
\end{aligned}
$$

η, m denote the step size and mini-batch size, respectively

s_i denote the eigenvalues of the sample covariance matrix

$\alpha, \gamma_y, \delta_y$ depend on n, p, m, η, s_i, y, but not t (see paper for details)

Implications/observations

$t = 1/\lambda$ scaling...

Appears in Rosasco & Poggio's papers and lectures
The implicit regularization of GD, and the connection to ridge, is, in their presentations, a key toy example for "why GD works in deep learning"
- `youtube.com/watch?v=4yLCuZnhkdI&t=3982` ..."number of iterations becomes $1/\lambda$"

Implications/observations

Result(s) hold across the entire optimization path

No strong conditions placed on the data matrix X

Also, have the following lower bound under oracle tuning

$$\inf_{\lambda \geq 0} \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}_{\lambda}; \beta_*) \leq \inf_{t \geq 0} \mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}_{t}; \beta_*)$$

Implications/observations

Proof: stochastic calculus (Ito's rule) uses the special covariance structure of the diffusion coefficient $Q_\eta(\beta_t)$ for least squares

$$\mathrm{Risk}(\hat{\beta}_t^{\mathrm{sgf}}) = \mathrm{Bias}^2(\hat{\beta}_t^{\mathrm{sgf}}) + \mathrm{Var}_\varepsilon(\hat{\beta}_t^{\mathrm{gf}}) + \mathrm{tr}\,\mathbb{E}_\varepsilon[\mathrm{Cov}_I(\hat{\beta}_t^{\mathrm{sgf}})|\varepsilon].$$

$$\mathrm{tr}\,\mathrm{Cov}_I(\hat{\beta}_t^{\mathrm{sgf}}) = O\left(\int_0^t \mathrm{tr}\,Q_\eta(\hat{\beta}_t^{\mathrm{sgf}}) \cdot \mathrm{tr}\,\hat{\Sigma}\exp[2(\tau - t)\hat{\Sigma}]d\tau\right)$$

$$\mathrm{tr}\left(Q_\eta(\beta)\right) = O\left(F(\beta)\right)$$

Implications/observations

The second and third (variance) terms ...

- Depend on the signal-to-noise ratio; this is different from gradient flow (and linear smoothers in general, because stochastic gradient flow/descent are actually *randomized* linear smoothers)
- The second term decreases with time, just as a bias would (but is a variance); this is different from gradient flow (see lemma in the paper)

Implications/observations

The second and third (variance) terms ...

- Depend on the signal-to-noise ratio; this is different from gradient flow (and linear smoothers in general, because stochastic gradient flow/descent are actually *randomized* linear smoothers)
- The second term decreases with time, just as a bias would (but is a variance); this is different from gradient flow (see lemma in the paper)

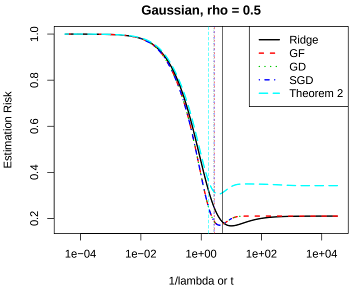
Third term vanishes (equals zero) if $p \geq n$ (can interpolate with overparametrization); "optimization variance smaller"

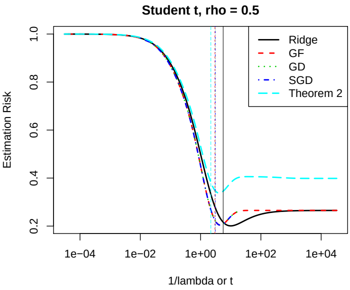
Outline

Synthetic data

Below, we show $n = 100, p = 10, m = 10$

- The bound (Theorem 2) tracks ridge's (and SGD's) risk(s) closely
- The bound / SGD achieve risk comparable to grad flow in less time
- See paper for other settings (e.g., high dimensions), coefficient error





Outline

Conclusion

Gave theoretical and empirical evidence showing stochastic gradient flow is closely related to ℓ_2 regularization - "why SGD works in statistical learning"

Interesting directions for future work

- Showing that stochastic gradient flow and SGD are, in fact, close
- Making the computational-statistical trade-off precise
- General convex losses
- Adaptive stochastic gradient methods

Thanks for helpful discussions: Misha Belkin, Quanquan Gu, J. Zico Kolter, Jason Lee, Yi-An Ma, Jascha Sohl-Dickstein, Daniel Soudry, and Matus Telgarsky. Part of this work was completed while ED was visiting the Simons Institute.

Thanks for listening!

References I

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. *arXiv preprint arXiv:1810.10082*, 2018.

Xiang Cheng, Peter L Bartlett, and Michael I Jordan. Quantitative w_1 convergence of langevin-like stochastic processes with non-convex potential state-dependent noise. *arXiv preprint arXiv:1907.03215*, 2019.

Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

Jianqing Fan, Wenyan Gong, Chris Junchi Li, and Qiang Sun. Statistical sparse online regression: A diffusion approximation perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 1017–1026, 2018.

References II

Yuanyuan Feng, Lei Li, and Jian-Guo Liu. Semi-groups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations. *arXiv preprint arXiv:1712.06509*, 2017.

Yuanyuan Feng, Tingran Gao, Lei Li, Jian-Guo Liu, and Yulong Lu. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *arXiv preprint arXiv:1902.00635*, 2019.

Stuart Geman and Chii-Ruey Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24(5):1031–1043, 1986.

David Gleich and Michael Mahoney. Anti-differentiating approximation algorithms: A case study with min-cuts, spectral, and flow. In *International Conference on Machine Learning*, pages 1018–1025, 2014.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

References III

Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv preprint arXiv:1705.07562*, 2017.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–40, 2019.

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

Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pages 2340–2348, 2016.

References IV

Michael W Mahoney. Approximate computation and implicit regularization for very large-scale data analysis. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, pages 143–154. ACM, 2012.

Michael W Mahoney and Lorenzo Orecchia. Implementing regularization implicitly via approximate eigenvector computation. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 121–128. Omnipress, 2011.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Continuous-time limit of stochastic gradient descent revisited. *NIPS-2015*, 2015.

Stephan Mandt, Matthew D Hoffman, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *The Journal of Machine Learning Research*, 18(1):4873–4907, 2017.

References V

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *arXiv preprint arXiv:1810.01075*, 2018.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. *arXiv preprint arXiv:1806.01796*, 2018.

Radford M Neal. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. *arXiv preprint arXiv:1802.08009*, 2018.

Tomaso Poggio, Andrzej Banburski, and Qianli Liao. Theoretical issues in deep networks: Approximation, optimization and generalization. *arXiv preprint arXiv:1908.09375*, 2019.

References VI

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

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

Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, pages 982–990, 2014.

Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45 (8):6056, 1992.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

References VII

Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pages 10608–10619, 2018.

Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

Yazhen Wang. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. *arXiv preprint arXiv:1711.09514*, 2017.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.