Krishnamurthy (Dj) Dvijotham

dvij@cs.washington.edu
http://dvij.github.io/

Objective

Research focused opportunities that leverage my scientific background.

- **Research interests:** Safe and Trustworthy AI, Formal Verification, Human-AI Collaboration, Uncertainty Qunatification, Machine Learning, Control Theory, and Optimization Theory.

Education

- **University of Washington, Seattle** — Seattle, WA, USA
  PhD in Computer Science and Engineering — Sep. 2008 – March 2014
  - Major: **Computer Science and Engineering**
- **Indian Institute of Technology, Bombay** — Mumbai, India
  Bachelor of Technology in Computer Science and Engineering — Jul. 2004 – May 2008
  - Major: **Computer Science and Engineering**

Work Experience

- **Staff Research Scientist, Google Brain** — Mountain View, CA, USA
  Google Brain — October 2021 - Present
  - Member of the Brain Privacy and Security team, co-leading several projects around Human-AI collaboration, interactive AI and formal verification of deep learning models with collaborators from several teams across Google Research, Google Health and DeepMind.
- **Staff Research Scientist, DeepMind** — London, UK
  DeepMind — Aug 2017 - Oct 2021
  - Founding member and co-lead of the Robust and Verified AI team working on formal verification, robustness, reliability and safety of deep learning models. Mangaged a team of 4 research scientists, mentored several interns and engineers. Delivered product impact on Android and Google Play Store, First author on several publications on formal verification and robustness of deep learning models.
- **Controls Engineer, Pacific Northwest National Laboratory** — Richland, WA, USA
  PNNL — Aug 2016 - Jul 2017
  - Led several projects on verification of safety properties of electric power grids under the Control of Complex Systems Initiative at PNNL.
- **Postdoctoral Fellow, California Institute of Technology** — Pasadena, CA, USA
  Caltech — Aug 2014 – Aug 2016
  - Research on control and optimization of electric power grids under high renewable energy penetration scenarios.

Research Awards

- **Best Paper Award, Conference on Uncertainty in Artificial Intelligence (UAI 2018)**
  August 2018
  - A Dual Approach to Scalable Verification of Deep Networks.
- **Best Paper Award, Conference on Constraint Programming (CP 2016)**
  September 2016
  - Universal Convexification via Risk Aversion.
- **Best Student Paper Award, Conference on Uncertainty in Artificial Intelligence (UAI 2014)**
  August 2014
  - Universal Convexification via Risk Aversion.
- **Best Student Paper Award, European Conference on Machine Learning (ECML 2008)**
  August 2008

- New closed-form upper bounds on the partition function.

Visa status

- Citizen of India, Permanent Resident of the USA

Service

- Area Chair, International Conference on Learning Representations (ICLR) 2019-2022
- Area Chair, Neural Information Processing Systems (NeurIPS) 2019-2022
- Action Editor, Transactions on Machine Learning Research (TMLR) 2022-Present
- Organizer, NIST Workshop on Assessing and Improving AI Trustworthiness: Current Contexts, Potential Paths (2021)
- Grant reviewer for Israeli Science Foundation (2022) and Sloan Foundation (2020)

Mentorship and Line Management Experience

- Line Manager to 4 research scientists at DeepMind from 2020-2021. Championed and secured promotions of two reports.
- Mentor to Several interns at DeepMind and Google Brain:
  - Rishav Chourasisa (Summer 2022)
  - Elizabeth Bondi (Summer 2021)
  - David Stutz (Summer 2021)
  - Sumanth Dathathri (Summer 2019)
  - Lily Weng (Summer 2019)
  - Johannes Welbl (Summer 2019)
  - Jamie Hayes (Summer 2019)
  - Rudy Bunel (Summer 2018)
  - Chenglong Wang (Summer 2018)
- Mentor to two interns (Ben Rapone and Haoxiang Yang) , and one postdoc (Thiagarajan Ramachandran) at PNNL

Publications

[1] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin Chadwick, Yoram Bachrach, Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of human-ai interaction in selective prediction. *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[2] Nicholas Carlini, Florian Tramer, Krishnamurthy, Dvijotham, and J. Zico Kolter. (certified!!) adversarial robustness for free!, 2022.

[3] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations*, 2022.

[4] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022.

[5] Harkirat Singh Behl, M Pawan Kumar, Philip Torr, and Krishnamurthy Dvijotham. Overcoming the convex barrier for simplex inputs. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] Leonard Berrada, Sumanth Dathathri, Krishnamurthy Dvijotham, Robert Stanforth, Rudy R Bunel, Jonathan Uesato, Sven Gowal, and M Pawan Kumar. Make sure you're unsure: A framework for verifying probabilistic specifications. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Harkirat Singh, M Pawan Kumar, Philip Torr, and Krishnamurthy Dj Dvijotham. Overcoming the convex barrier for simplex inputs. In *Advances in Neural Information Processing Systems*, 2021.

[8] Yu Weng, Suhyoun Yu, Krishnamurthy Dvijotham, and Hung Nguyen. Fixed-point theorem-based voltage stability margin estimation techniques for distribution systems with renewables. *IEEE Transactions on Industrial Informatics*, 2021.

[9] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis of robustness to distribution shifts. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[10] Haoxiang Yang, David P Morton, Chaithanya Bandi, and Krishnamurthy Dvijotham. Robust optimization for electricity generation. *INFORMS Journal on Computing*, 33(1):336–351, 2021.

[11] Navid Azizan, Yu Su, Krishnamurthy Dvijotham, and Adam Wierman. Optimal pricing in markets with nonconvex costs. *Operations Research*, 68(2):480–496, 2020.

[12] Rudy Bunel, Alessandro De Palma, Alban Desmaison, Krishnamurthy Dvijotham, Pushmeet Kohli, Philip Torr, and M Pawan Kumar. Lagrangian decomposition for neural network verification. In *Conference on Uncertainty in Artificial Intelligence*, pages 370–379. PMLR, 2020.

[13] Rudy R Bunel, Oliver Hinder, Srinadh Bhojanapalli, and Krishnamurthy Dvijotham. An efficient nonconvex reformulation of stagewise convex optimization problems. *Advances in Neural Information Processing Systems*, 33:8247–8258, 2020.

[14] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, and Pushmeet Kohli. The autoencoding variational autoencoder. *Advances in Neural Information Processing Systems*, 33:15077–15087, 2020.

[15] Sumanth Dathathri, Krishnamurthy Dvijotham, Alexey Kurakin, Aditi Raghunathan, Jonathan Uesato, Rudy R Bunel, Shreya Shankar, Jacob Steinhardt, Ian Goodfellow, Percy S Liang, et al. Enabling certification of verification-agnostic networks via memory-efficient semidefinite programming. *Advances in Neural Information Processing Systems*, 33:5318–5331, 2020.

[16] Krishnamurthy Dvijotham, Yuval Rabani, and Leonard J Schulman. Convergence of incentive-driven dynamics in fisher markets. *Games and Economic Behavior*, 2020.

[17] Krishnamurthy Dj Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020.

[18] Krishnamurthy Dj Dvijotham, Robert Stanforth, Sven Gowal, Chongli Qin, Soham De, and Pushmeet Kohli. Efficient neural network verification with exactness characterization. In *Uncertainty in artificial intelligence*, pages 497–507. PMLR, 2020.

[19] Sven Gowal, Chongli Qin, Po-Sen Huang, Taylan Cemgil, Krishnamurthy Dvijotham, Timothy Mann, and Pushmeet Kohli. Achieving robustness in the wild via adversarial mixing with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1211–1220, 2020.

[20] Johannes Welbl, Po-Sen Huang, Robert Stanforth, Sven Gowal, Krishnamurthy Dj Dvijotham, Martin Szummer, and Pushmeet Kohli. Towards verified robustness under text deletion interventions. In *International Conference on Learning Representations*, 2020.

[21] Anton Zhernov, Krishnamurthy Dj Dvijotham, Ivan Lobov, Dan A Calian, Michelle Gong, Natarajan Chandrashekar, and Timothy A Mann. The nodehopper: Enabling low latency ranking with constraints via a fast dual solver. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1285–1294, 2020.

[22] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dj Dvijotham, and Pushmeet Kohli. Adversarially robust representations with smooth encoders. In *International Conference on Learning Representations*, 2019.

[23] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Timothy A Mann, and Pushmeet Kohli. A dual approach to verify and train deep networks. In *IJCAI*, pages 6156–6160, 2019.

[24] Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4842–4851, 2019.

[25] Po-Sen Huang, Robert Stanforth, Johannes Welbl, Chris Dyer, Dani Yogatama, Sven Gowal, Krishnamurthy Dvijotham, and Pushmeet Kohli. Achieving verified robustness to symbol substitutions via interval bound propagation. *ACL 2019*, 2019.

[26] Dongchan Lee, Hung D Nguyen, Krishnamurthy Dvijotham, and Konstantin Turitsyn. Convex restriction of power flow feasibility sets. *IEEE Transactions on Control of Network Systems*, 6(3):1235–1245, 2019.

[27] Parikshit Pareek, Konstantin Turitsyn, Krishnamurthy Dvijotham, and Hung D Nguyen. A sufficient condition for small-signal stability and construction of robust stability region. In *2019 IEEE Power & Energy Society General Meeting (PESGM)*, pages 1–5. IEEE, 2019.

[28] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Chongli Qin, Brendan O'Donoghue, Rudy Bunel, Robert Stanforth, Sven Gowal, Jonathan Uesato, Grzegorz Swirszcz, Pushmeet Kohli, et al. Verification of non-linear specifications for neural networks. *International Conference on Learning Representations (ICLR) 2019*, 2019.

[30] Chenglong Wang, Rudy Bunel, Krishnamurthy Dvijotham, Po-Sen Huang, Edward Grefenstette, and Pushmeet Kohli. Knowing when to stop: Evaluation and verification of conformity to output-size specifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12260–12269, 2019.

[31] Tsui-Wei Weng, Krishnamurthy Dj Dvijotham, Jonathan Uesato, Kai Xiao, Sven Gowal, Robert Stanforth, and Pushmeet Kohli. Toward evaluating robustness of deep reinforcement learning with continuous control. In *International Conference on Learning Representations*, 2019.

[32] S. Misra, D. K. Molzahn, and K. Dvijotham. Optimal adaptive linearizations of the ac power flow equations. In *2018 Power Systems Computation Conference (PSCC)*, pages 1–7, June 2018.

[33] Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. In *Proceedings of the Thirty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, pages 162–171, Corvallis, Oregon, 2018. AUAI Press.

[34] Vinod Nair, Krishnamurthy Dvijotham, Iain Dunning, and Oriyol Vinyals. Learning fast optimizers for contextual stochastic integer programs. In *Proceedings of the Thirty-Fourth Conference on*

Uncertainty in Artificial Intelligence, UAI 2018, Monterrey, California, USA, August 6-9, 2018, 2018.

[35] H. D. Nguyen, K. Dvijotham, and K. Turitsyn. Constructing convex inner approximations of steady-state security regions. *IEEE Transactions on Power Systems*, pages 1–1, 2018.

[36] H. D. Nguyen, K. Dvijotham, S. Yu, and K. Turitsyn. A framework for robust long-term voltage stability of distribution systems. *IEEE Transactions on Smart Grid*, pages 1–1, 2018.

[37] K. Dvijotham, E. Mallada, and J. W. Simpson-Porco. High-voltage solution in radial power networks: Existence, properties, and equivalent algorithms. *IEEE Control Systems Letters*, 1(2):322–327, Oct 2017.

[38] K. Dvijotham, H. Nguyen, and K. Turitsyn. Solvability regions of affinely parameterized quadratic equations. *IEEE Control Systems Letters*, PP(99):1–1, 2017.

[39] Krishnamurthy Dvijotham, Yuval Rabani, and Leonard Schulman. Convergence of incentive-driven dynamics in fisher markets. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, January 10-12, 2017*, pages 2039–2052, 2017.

[40] N. Azizan Ruhi, K. Dvijotham, N. Chen, and A. Wierman. Opportunities for price manipulation by aggregators in electricity markets. *IEEE Transactions on Smart Grid*, PP(99):1–1, 2017.

[41] Y. Tang, K. Dvijotham, and S. Low. Real-time optimal power flow. *IEEE Transactions on Smart Grid*, PP(99):1–1, 2017.

[42] D. Wu, D. K. Molzahn, B. C. Lesieutre, and K. Dvijotham. A deterministic method to identify multiple local extrema for the ac optimal power flow problem. *IEEE Transactions on Power Systems*, PP(99):1–1, 2017.

[43] K. Dvijotham and D. Molzahn. Error bounds on the dc power flow approximation: A convex relaxation approach. In *2016 55th IEEE Conference on Decision and Control (CDC)*, pages 23–30, Dec 2016.

[44] Krishnamurthy Dvijotham, Michael Chertkov, Pascal Van Hentenryck, Marc Vuffray, and Sidhant Misra. Graphical models for optimal power flow. *Constraints*, pages 1–26, 2016.

[45] K. Dvijotham, M. Chertkov, and S. Low. A differential analysis of the power flow equations. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 23–30, Dec 2015.

[46] K. Dvijotham and M. Chertkov. Convexity of structure preserving energy functions in power transmission: Novel results and applications. In *2015 American Control Conference (ACC)*, pages 5035–5042, July 2015.

[47] Krishnamurthy Dvijotham. Systems of quadratic equations: Efficient solution algorithms and conditions for solvability. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1027–1031. IEEE, 2015.

[48] Krishnamurthy Dvijotham, Emanuel Todorov, and Maryam Fazel. Convex structured controller design in finite horizon. *IEEE Transactions on Control of Network Systems*, 2(1):1–10, 2015.

[49] Krishnamurthy Dvijotham. *Automating Stochastic Optimal Control*. PhD thesis, 2014.

[50] Krishnamurthy Dvijotham, Misha Chertkov, and Scott Backhaus. Storage sizing and placement through operational and uncertainty-aware simulations. In *2014 47th Hawaii International Conference on System Sciences*, pages 2408–2416. IEEE, 2014.

[51] Krishnamurthy Dvijotham, Maryam Fazel, and Emanuel Todorov. Convex risk averse control design. In *53rd IEEE Conference on Decision and Control*, pages 4020–4025. IEEE, 2014.

[52] Krishnamurthy Dvijotham, Maryam Fazel, and Emanuel Todorov. Universal convexification via risk-aversion. In *Proceedings of the Thirtieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-14)*, pages 162–171, Corvallis, Oregon, 2014. AUAI Press.

[53] K. Dvijotham and R. Sharma. Battery life estimation in a real-time energy management system. In *2013 IEEE Power Energy Society General Meeting*, pages 1–5, July 2013.

[54] K. Dvijotham and E. Todorov. *Linearly Solvable Optimal Control*, pages 119–141. John Wiley & Sons, Inc., 2013.

[55] Krishnamurthy Dvijotham, Evangelos Theodorou, Emanuel Todorov, and Maryam Fazel. Convexity of optimal linear controller design. In *52nd IEEE Conference on Decision and Control*. IEEE, 2013.

[56] Krishnamurthy Dvijotham, Emanuel Todorov, and Maryam Fazel. Convex control design via covariance minimization. In *Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2013.

[57] Evangelos Theodorou, Krishnamurthy Dvijotham, and Emo Todorov. Time varying nonlinear policy gradients. In *CDC*, pages 7765–7770, 2013.

[58] Evangelos Theodorou, D Krishnamurthy, and Emo Todorov. From information theoretic dualities to path integral and kullback-leibler control: Continuous and discrete time formulations. In *The Sixteenth Yale Workshop on Adaptive and Learning Systems*, 2013.

[59] Krishnamurthy Dvijotham, Scott Backhaus, and Michael Chertkov. Distributed control of generation in a transmission grid with a high penetration of renewables. In *Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference on*, pages 635–640. IEEE, 2012.

[60] Krishnamurthy Dvijotham and Emo Todorov. Linearly solvable markov games. In *2012 American Control Conference (ACC)*, pages 1845–1850. IEEE, 2012.

[61] Krishnamurthy Dvijotham and Emanuel Todorov. A unifying framework for linearly solvable control. In *Proceedings of the Twenty-Seventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-11)*, pages 179–186, Corvallis, Oregon, 2011. AUAI Press.

[62] Krishnamurthy Dvijotham and Maryam Fazel. A nullspace analysis of the nuclear norm heuristic for rank minimization. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3586–3589. IEEE, 2010.

[63] Krishnamurthy Dvijotham and Emanuel Todorov. Inverse optimal control with linearly-solvable mdps. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 335–342, 2010.

[64] Krishnamurthy Dvijotham, Soumen Chakrabarti, and Subhasis Chaudhuri. New closed-form bounds on the partition function. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 8–8. Springer Berlin Heidelberg, 2008.