CS480/680: Introduction to Machine Learning

Lec 13: Generative Adversarial Networks

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

March 12, 2024

Recap: MLE

- Given training data $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \sim q(\mathbf{x})$, the data density
- Parameterize $p_{\boldsymbol{\theta}}(\mathbf{x})$, the model density, e.g., Gaussian mixture
- Estimate $\boldsymbol{\theta}$ by minimizing some "distance" between q (the unknown data density) and $p_{\boldsymbol{\theta}}$ (the chosen model density):

$$\min_{\boldsymbol{\theta}} \ \mathsf{KL}(q \| p_{\boldsymbol{\theta}}) \quad \equiv \quad \int -\log p_{\boldsymbol{\theta}}(\mathbf{x}) \cdot q(\mathbf{x}) \, \mathrm{d}\mathbf{x} \quad \approx \quad -\frac{1}{n} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i)$$

- After training, can generate new data $\mathsf{X} \sim p_{\boldsymbol{\theta}}(\mathbf{x})$
- Need a training sample $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from q and an explicit form of $p_{\boldsymbol{\theta}}$

Generative Adversarial Networks

$$\min_{\boldsymbol{\theta}} \ \mathsf{KL}(q \| p_{\boldsymbol{\theta}}) \quad \equiv \quad \int - \log p_{\boldsymbol{\theta}}(\mathbf{x}) \cdot q(\mathbf{x}) \, \mathrm{d}\mathbf{x} \quad \approx \quad -\frac{1}{n} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i)$$

- What if we do not have an explicit form of $p_{\boldsymbol{\theta}}$?
- Would a sample from $p_{\boldsymbol{\theta}}$ suffice?

$$p(\mathbf{x}) = (2\pi)^{-d/2}[\det(S)]^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top S^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

- Draw $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathrm{Id})$
- Set $\boxed{\mathbf{x} = L\mathbf{n} + \boldsymbol{\mu}}$
- Recall: linear transformation of Gaussian is Gaussian
  - $\mathbb{E}(\mathbf{x}) = \mathbb{E}(L\mathbf{n} + \boldsymbol{\mu}) = \boldsymbol{\mu}$
  - $\mathbb{E}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top = L \cdot \mathbb{E}(\mathbf{n}\mathbf{n}^\top) \cdot L^\top = LL^\top =: S$
- Let $p_{\boldsymbol{\theta}}(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, S)$ and we have found a way to sample from $p_{\boldsymbol{\theta}}$

Push-forward Maps

Theorem: Representation through push-forward

Let r be any continuous distribution on $\mathbb{R}^h$. For any distribution p on $\mathbb{R}^d$, there exist push-forward maps $\mathbf{T} : \mathbb{R}^h \to \mathbb{R}^d$ such that

$$\mathsf{Z} \sim r \implies \mathbf{T}(\mathsf{Z}) \sim p$$

- Result is false if r has a delta mass at some point
- W.l.o.g. we may simply take r to be standard Gaussian noise
- Mapping $\mathbf{T}$ is not unique; can add additional restrictions to induce uniqueness
- Mapping $\mathbf{T}$ can be weird if $h \ll d$...

A Whole New World

- Given training data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \sim q(\mathbf{x})$, the data density
- Can now parameterize the model density $p_{\boldsymbol{\theta}}(\mathbf{x})$ as the push-forward of standard Gaussian, i.e., the density of $\mathbf{T}_\theta(\mathsf{Z})$, where $\mathsf{Z} \sim \mathcal{N}(\mathbf{0}, \mathrm{Id})$ is noise
- Estimate $\boldsymbol{\theta}$ by minimizing some "distance" between q (the unknown data density) and $p_{\boldsymbol{\theta}}$ (the chosen model density):

$$\min_{\boldsymbol{\theta}} \ \mathsf{KL}(q \| p_{\boldsymbol{\theta}}) \equiv \int -\log p_{\boldsymbol{\theta}}(\mathbf{x}) \cdot q(\mathbf{x}) \, \mathrm{d}\mathbf{x} \approx -\frac{1}{n} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i)$$

- After training, can generate new data $\mathsf{X} = \mathbf{T}_{\boldsymbol{\theta}}(\mathsf{Z})$ where $\mathsf{Z} \sim \mathcal{N}(\mathbf{0}, \mathrm{Id})$
- Do not have the density $p_{\boldsymbol{\theta}}$; can only sample from it...

Minimizing KL with Both Inputs UNKNOWN...

$$\min_{\boldsymbol{\theta}} \ \mathrm{KL}(q \| p_{\boldsymbol{\theta}}) = \int \log \frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \cdot q(\mathbf{x}) \, \mathrm{d}\mathbf{x} \equiv \int \underbrace{\frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \left[\log \frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} - 1 \right]}_{f\left(\frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right)} \cdot p_{\boldsymbol{\theta}}(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

- $f : \mathbb{R}_+ \to \mathbb{R}, \quad t \mapsto t \log t - t + 1$ is a convex function

Definition: Fenchel conjugate

The conjugate of any function f is the convex function $f^*(s) := \max_t \ st - f(t)$.
Moreover, if f is convex and continuous, then $f = f^{**}$.

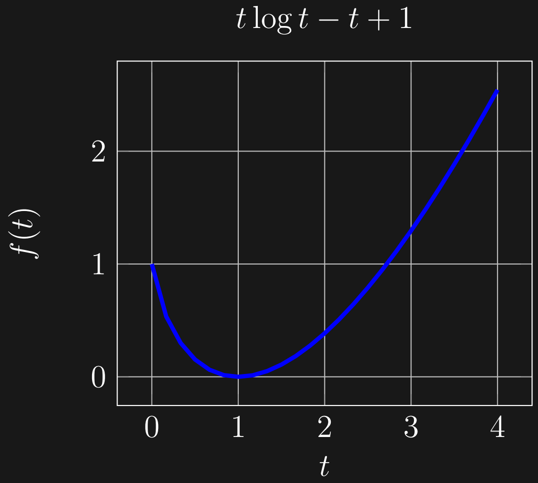
$$f^*(s) = \left[\max_t st - f(t)\right] = \max_t st - t\log t + t - 1$$

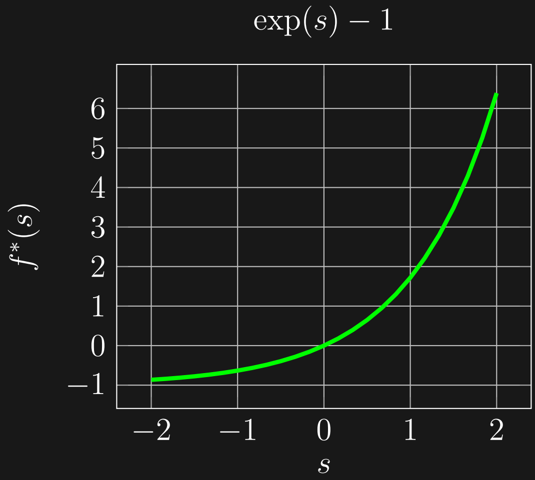
- Setting derivative to 0 we obtain $s = \log t$, i.e., $t = \exp(s)$
- Plugging back in we obtain

$$f^*(s) = \exp(s) - 1$$

- Since f is convex and continuous, we have

$$f(t) = \left[\max_s \; st - f^*(s)\right] = \left[\max_s \; st - \exp(s) + 1\right]$$


$t \log t - t + 1$
$f(t)$
t
0
1
2
0
1
2
3
4


$\exp(s) - 1$
$f^*(s)$
s
6
5
4
3
2
1
0
−1
−2
−1
0
1
2

Duality

$$\min_{\boldsymbol{\theta}} \ \mathsf{KL}(q \| p_{\boldsymbol{\theta}}) \equiv \int \underbrace{\frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\left[\log \frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} - 1\right]}_{f\left(\frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right) \ -1} \cdot p_{\boldsymbol{\theta}}(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

$$\equiv \int \left[\max_s \ s\frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} - \exp(s)\right] \cdot p_{\boldsymbol{\theta}}(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

$$= \int \left[\max_s \ sq(\mathbf{x}) - \exp(s) p_{\boldsymbol{\theta}}(\mathbf{x})\right] \mathrm{d}\mathbf{x}$$

$$= \max_{S:\mathbb{R}^d \to \mathbb{R}} \int \left[S(\mathbf{x})q(\mathbf{x}) - \exp(S(\mathbf{x}))p_{\boldsymbol{\theta}}(\mathbf{x})\right] \mathrm{d}\mathbf{x}$$

$$\approx \max_{S:\mathbb{R}^d \to \mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} S(\mathbf{x}_i) - \frac{1}{m}\sum_{j=1}^{m} \exp\left[S(\mathbf{T}_{\boldsymbol{\theta}}(\mathbf{z}_j))\right]$$

Putting Things Together

$$\min_{\boldsymbol{\theta}} \ \mathsf{KL}(q \| p_{\boldsymbol{\theta}}) \quad \approx \quad \min_{\boldsymbol{\theta}} \ \max_{\boldsymbol{\phi}} \ \frac{1}{n}\sum_{i=1}^{n} S_{\boldsymbol{\phi}}(\mathbf{x}_i) - \frac{1}{m}\sum_{j=1}^{m} \exp\left[S_{\boldsymbol{\phi}}(\mathbf{T}_{\boldsymbol{\theta}}(\mathbf{z}_j))\right]$$

- $\mathbf{T}_{\boldsymbol{\theta}}$: generator, mapping latent noise $\mathbf{z}$ to observation $\mathbf{x}$
- $S_{\boldsymbol{\phi}}$: discriminator, distinguishing data $\mathbf{x}$ from generation $\mathbf{T}_{\boldsymbol{\theta}}(\mathbf{z})$
- Both are neural networks parameterized by weights $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, resp.
- A minimax game between generator and discriminator
- At equilibrium: generator learns data and discriminator cannot distinguish

Support Considerations

$$\min_{\boldsymbol{\theta}} \ \mathsf{KL}(q \| p_{\boldsymbol{\theta}}) = \int \log \frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \cdot q(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

- What happens if $p_{\boldsymbol{\theta}}(\mathbf{x}) \approx 0$?

$$\min_{\boldsymbol{\theta}} \ \mathsf{KL}(p_{\boldsymbol{\theta}} \| q) = \int \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x})}{q(\mathbf{x})} \cdot p_{\boldsymbol{\theta}}(\mathbf{x}) \, \mathrm{d}\mathbf{x}$$

- What happens if $p_{\boldsymbol{\theta}}(\mathbf{x}) \approx 0$?

$$\min_{\boldsymbol{\theta}} \ \mathsf{KL}(q \| p_{\boldsymbol{\theta}}) + \mathsf{KL}(p_{\boldsymbol{\theta}} \| q) = \int \left[\log \frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \cdot q(\mathbf{x}) + \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x})}{q(\mathbf{x})} \cdot p_{\boldsymbol{\theta}}(\mathbf{x}) \right] \mathrm{d}\mathbf{x}$$

Generative Adversarial Networks

$$\boxed{\mathsf{JS}(q \| p_{\boldsymbol{\theta}}) := \mathsf{KL}(q \| \tfrac{q+p_{\boldsymbol{\theta}}}{2}) + \mathsf{KL}(p_{\boldsymbol{\theta}} \| \tfrac{q+p_{\boldsymbol{\theta}}}{2})}$$

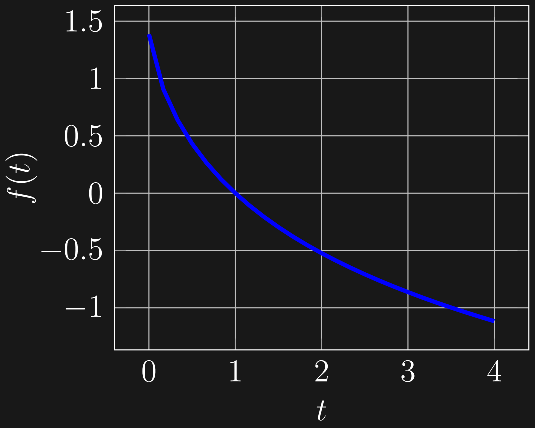
$$= \int q(\mathbf{x}) \log \frac{2q(\mathbf{x})}{q(\mathbf{x}) + p_{\boldsymbol{\theta}}(\mathbf{x})} + p_{\boldsymbol{\theta}}(\mathbf{x}) \log \frac{2p_{\boldsymbol{\theta}}(\mathbf{x})}{q(\mathbf{x}) + p_{\boldsymbol{\theta}}(\mathbf{x})} \,\mathrm{d}\mathbf{x}$$

$$= \int \underbrace{\left[\frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})} \log \frac{q(\mathbf{x})/p_{\boldsymbol{\theta}}(\mathbf{x})}{q(\mathbf{x})/p_{\boldsymbol{\theta}}(\mathbf{x}) + 1} + \log \frac{1}{q(\mathbf{x})/p_{\boldsymbol{\theta}}(\mathbf{x}) + 1} + \log 4 \right]}_{f\left(\frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right)} \cdot p_{\boldsymbol{\theta}}(\mathbf{x}) \,\mathrm{d}\mathbf{x}$$

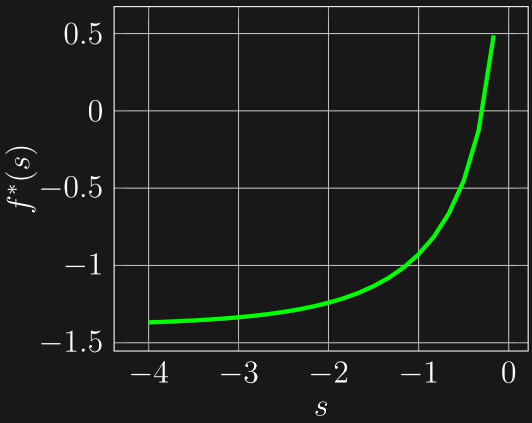
- $f : \mathbb{R}_+ \to \mathbb{R}, \quad t \mapsto t \log t - (t+1)\log(t+1) + \log 4$ is convex
- $f^*(s) = -\log(1 - \exp(s)) - \log 4$

I. Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. 2014.

$$t\log t - (t+1)\log(t+1) + \log 4$$



$$-\log(1-\exp(s)) - \log 4$$



$$\min_{\boldsymbol{\theta}} \ \texttt{JS}(q\|p_{\boldsymbol{\theta}}) \ \approx \ \min_{\boldsymbol{\theta}} \ \max_{\boldsymbol{\phi}} \ \frac{1}{n}\sum_{i=1}^{n} S_{\boldsymbol{\phi}}(\mathbf{x}_i) + \frac{1}{m}\sum_{j=1}^{m} \log\left[1 - \exp(S_{\boldsymbol{\phi}}(\mathbf{T}_{\boldsymbol{\theta}}(\mathbf{z}_j)))\right] - \log 4$$

- Apply change of variable: $S_{\boldsymbol{\phi}} \leftarrow \log S_{\boldsymbol{\phi}}$:

$$\min_{\boldsymbol{\theta}} \ \texttt{JS}(q\|p_{\boldsymbol{\theta}}) \ \approx \ \min_{\boldsymbol{\theta}} \ \max_{\boldsymbol{\phi}} \ \frac{1}{n}\sum_{i=1}^{n} \log S_{\boldsymbol{\phi}}(\mathbf{x}_i) + \frac{1}{m}\sum_{j=1}^{m} \log\left[1 - S_{\boldsymbol{\phi}}(\mathbf{T}_{\boldsymbol{\theta}}(\mathbf{z}_j))\right]$$

- Let $\mathsf{y}(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \text{ is real} \\ 0, & \text{if } \mathbf{x} \in \{\mathbf{T}_{\boldsymbol{\theta}}(\mathbf{z}_1), \ldots, \mathbf{T}_{\boldsymbol{\theta}}(\mathbf{z}_m)\} \text{ is generated} \end{cases}$
- Let $\mathfrak{p}_1(\mathbf{x}) := S_{\boldsymbol{\phi}}(\mathbf{x}) \in [0,1]$ be the probability of $\mathbf{x}$ being real, and $\mathfrak{p}_0 = 1 - \mathfrak{p}_1$:

$$\min_{\boldsymbol{\theta}} \ \underbrace{\max_{\boldsymbol{\phi}} \ \hat{\mathbb{E}}_{\mathsf{X}} \log \mathfrak{p}_{\mathsf{y}(\mathsf{X})}}_{\text{logistic regression}}$$


Z
T_{θ^*}
X

https://www.whichfaceisreal.com/

https://thispersondoesnotexist.com/

T. Karras et al. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4396–4405, T. Karras et al. "Analyzing and Improving the Image Quality of StyleGAN". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8107–8116, T. Karras et al. "Alias-Free Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems 34*. 2021.

More Generally

$$\min_{\boldsymbol{\theta}} \boxed{\mathbb{D}_f(q \| p_{\boldsymbol{\theta}}) := \int f\left(\frac{q(\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{x})}\right) p_{\boldsymbol{\theta}}(\mathbf{x})\,\mathrm{d}\mathbf{x}}$$

$$\approx \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \frac{1}{n}\sum_{i=1}^{n} S_{\boldsymbol{\phi}}(\mathbf{x}_i) - \frac{1}{m}\sum_{j=1}^{m} f^*\left[S_{\boldsymbol{\phi}}\big(\mathbf{T}_{\boldsymbol{\theta}}(\mathbf{z}_j)\big)\right]$$

- Choose any (strictly) convex function $f : \mathbb{R}_+ \to \mathbb{R}$, with normalization $f(1) = 0$
- The f-divergence $\mathbb{D}_f$ such that $\mathbb{D}_f(q\|p) \geq 0$, with equality iff $q = p$
  - $f(t) = t \log t - t + 1$ gives KL
  - $f(t) = -\log t$ gives reverse KL
  - $f(t) = t \log t - (t+1)\log(t+1) + \log 4$ gives JS

I. Csiszár. "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten". *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének közleményei*, vol. 8 (1963), pp. 85–108, S. Nowozin et al. "f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization". In: *Advances in Neural Information Processing Systems*. 2016.

Other Variants

- MMD-GAN. Discriminator from the unit ball of an RKHS:

$$\min_{\boldsymbol{\theta}} \max_{S \in \mathcal{H}_k, \|S\| \leq 1} \mathbb{E}_{\mathsf{X} \sim q,\ \mathsf{Z} \sim r} [S(\mathsf{X}) - S(\mathbf{T}_{\boldsymbol{\theta}}(\mathsf{Z}))]^2$$

  - reproducing property: $S(\mathsf{X}) - S(\mathbf{T}_{\boldsymbol{\theta}}(\mathsf{Z})) = \langle k(\cdot, \mathsf{X}) - k(\cdot, \mathbf{T}_{\boldsymbol{\theta}}(\mathsf{Z})), S \rangle$
  - apply Cauchy-Schwarz to eliminate S

- Wasserstein GAN. Discriminator from all Lipschitz continuous functions:

$$\min_{\boldsymbol{\theta}} \underbrace{\max_{\|S\|_{\mathrm{Lip}} \leq 1} \mathbb{E}_{\mathsf{X} \sim q,\ \mathsf{Z} \sim r} [S(\mathsf{X}) - S(\mathbf{T}_{\boldsymbol{\theta}}(\mathsf{Z}))]}_{\text{Wasserstein distance}}$$

  - parameterize S as a neural network

G. K. Dziugaite et al. "Training generative neural networks via maximum mean discrepancy optimization". In: *Conference on Uncertainty in Artificial Intelligence*. 2015, Y. Li et al. "Generative Moment Matching Networks". In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015, M. Arjovsky et al. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. 2017.