Information theoretic perspectives on learning algorithms

Varun Jog

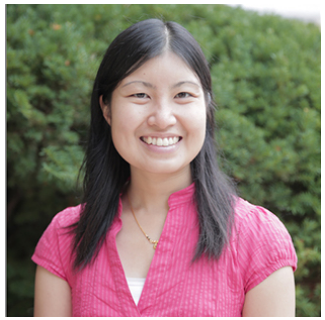
University of Wisconsin - Madison
Electrical and Computer Engineering

R Narasimhan Memorial Lecture

January 2, 2019

Work with Adrian Tovar-Lopez (Math), Ankit Pensia (CS), Po-Ling Loh (Stats)





Curve fitting

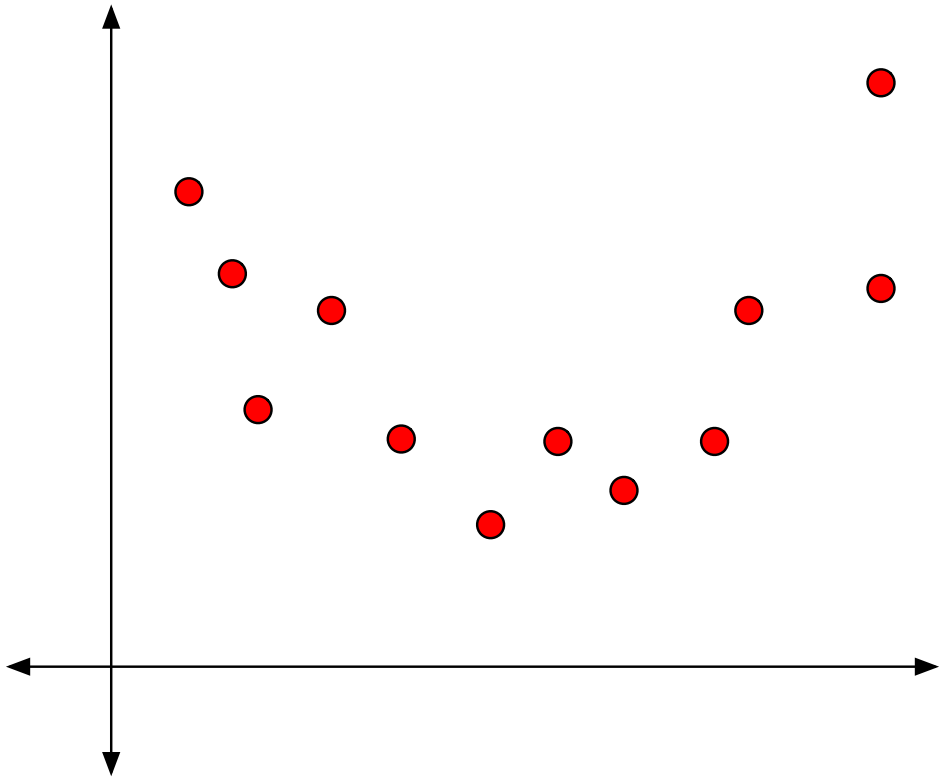
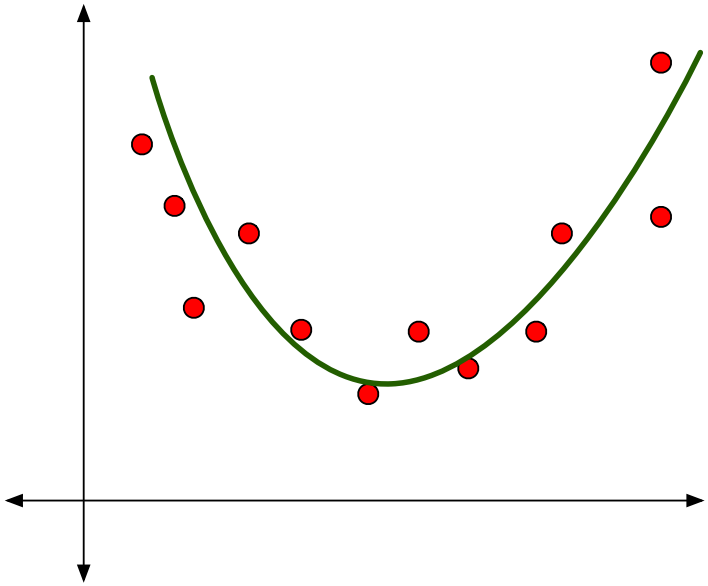
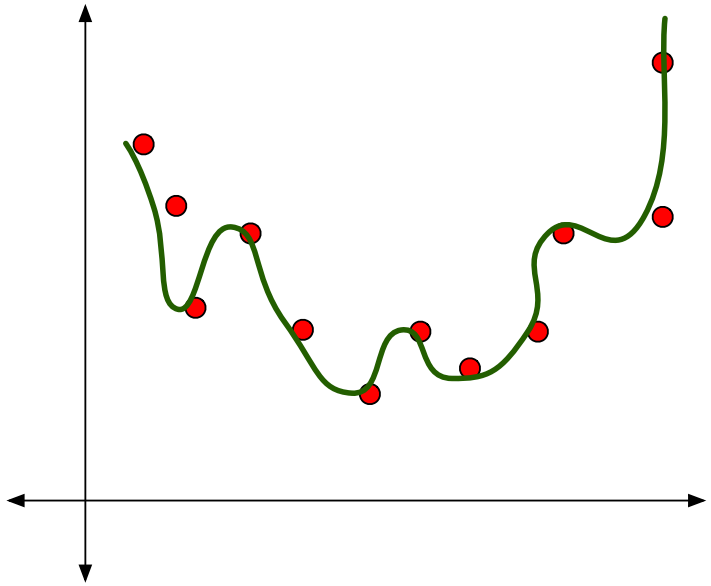


Figure: Given N points in $\mathbb{R}^2$, fit a curve

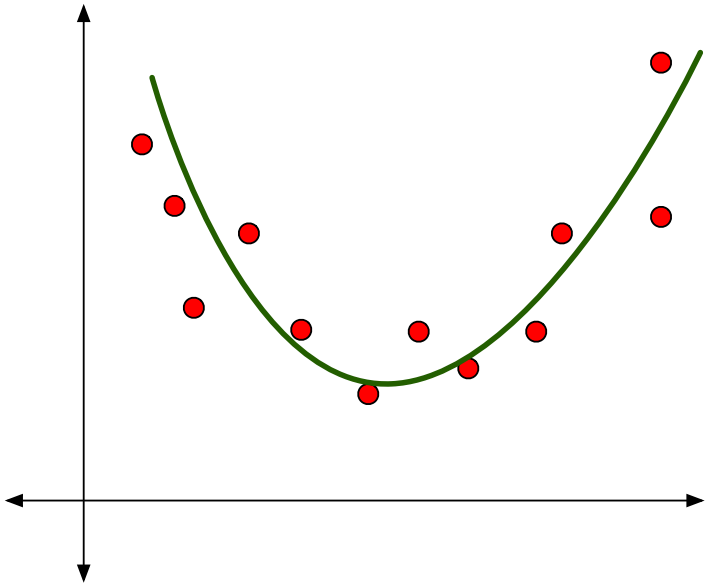
- Forward problem: From dataset to curve

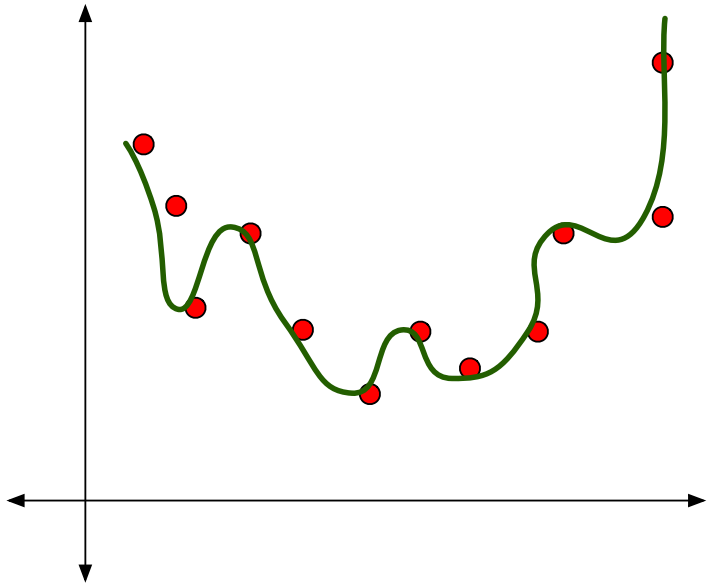
Finding the right "fit"





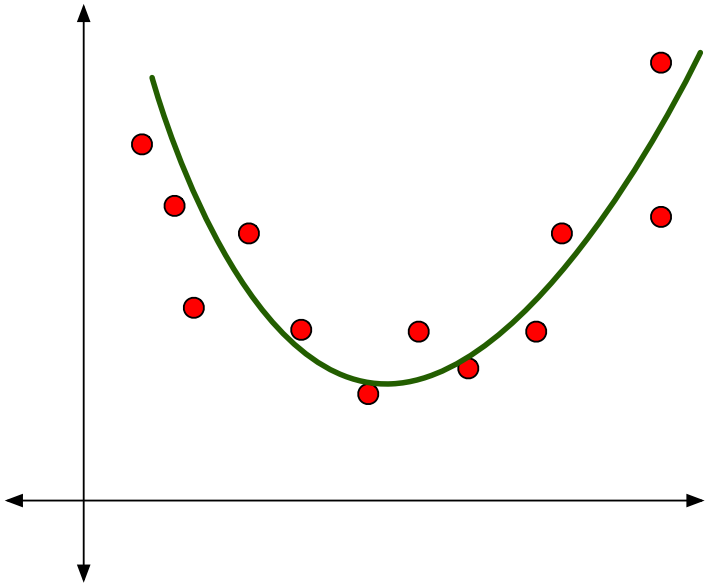
Finding the right "fit"

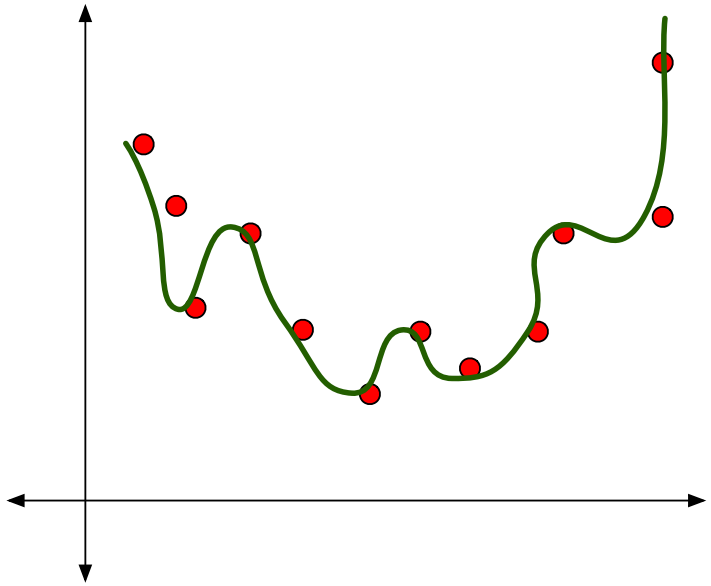




- Left is fit, right is overfit

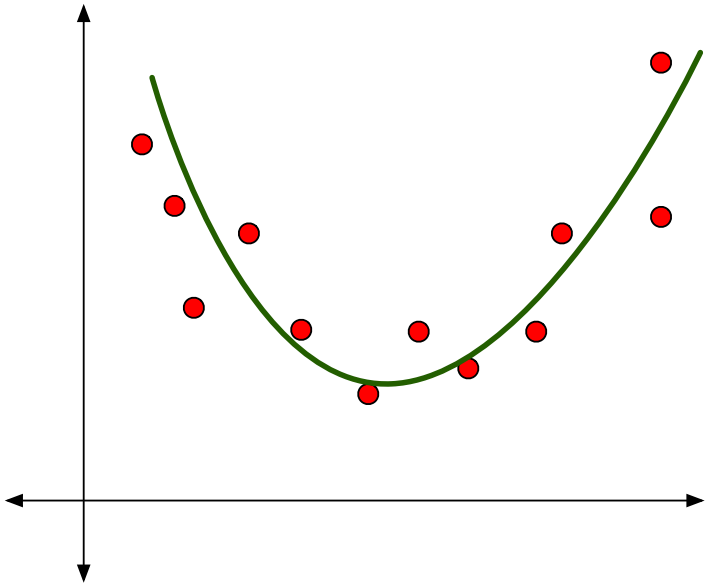
Finding the right "fit"

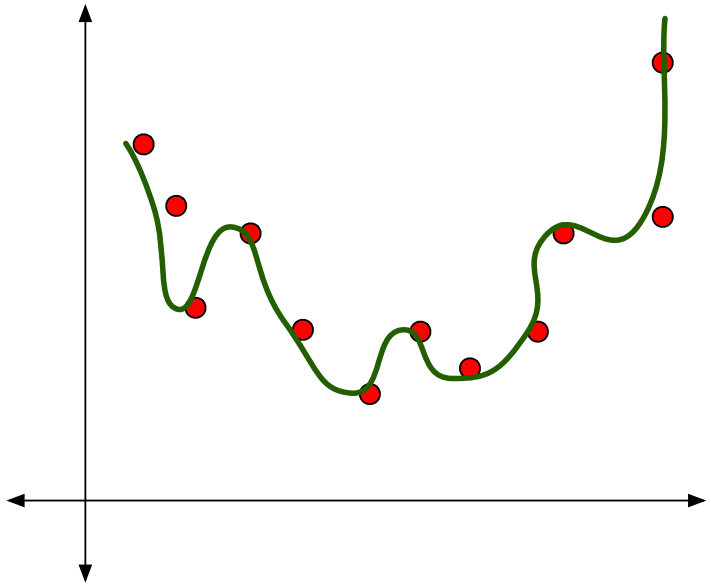




- Left is fit, right is overfit
- Too wiggly

Finding the right "fit"





- Left is fit, right is overfit
- Too wiggly
- Not stable

Guessing points from curve

Guessing points from curve

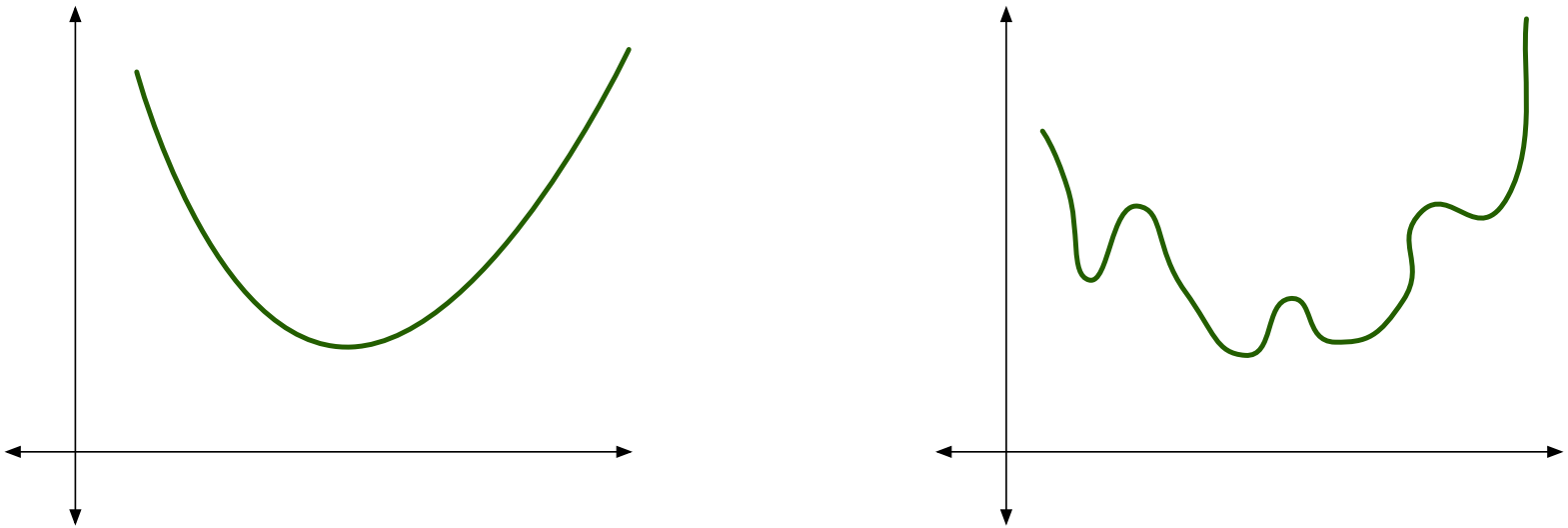


Figure: Given curve, find N points

- Backward problem: From curve to dataset

Guessing points from curve

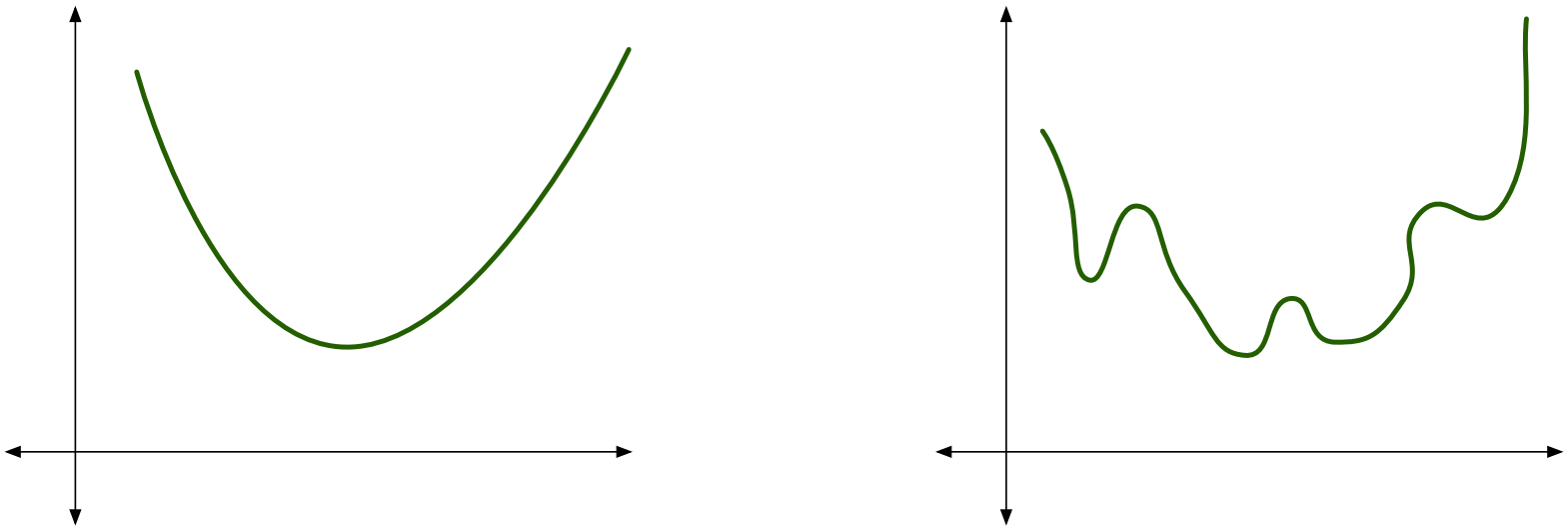


Figure: Given curve, find N points

- Backward problem: From curve to dataset
- Backward problem easier for overfitted curve!

Guessing points from curve

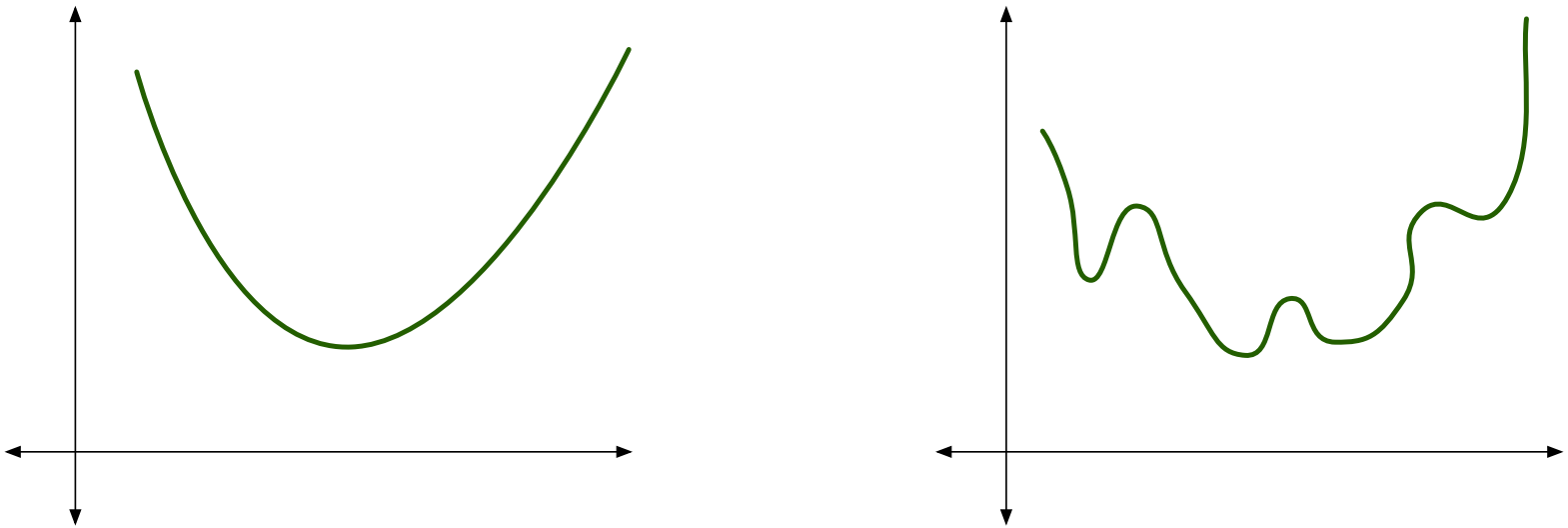


Figure: Given curve, find N points

- Backward problem: From curve to dataset
- Backward problem easier for overfitted curve!
- Curve contains more information about dataset

This talk

This talk

- Explore information and overfitting connection (Xu & Raginsky, 2017)

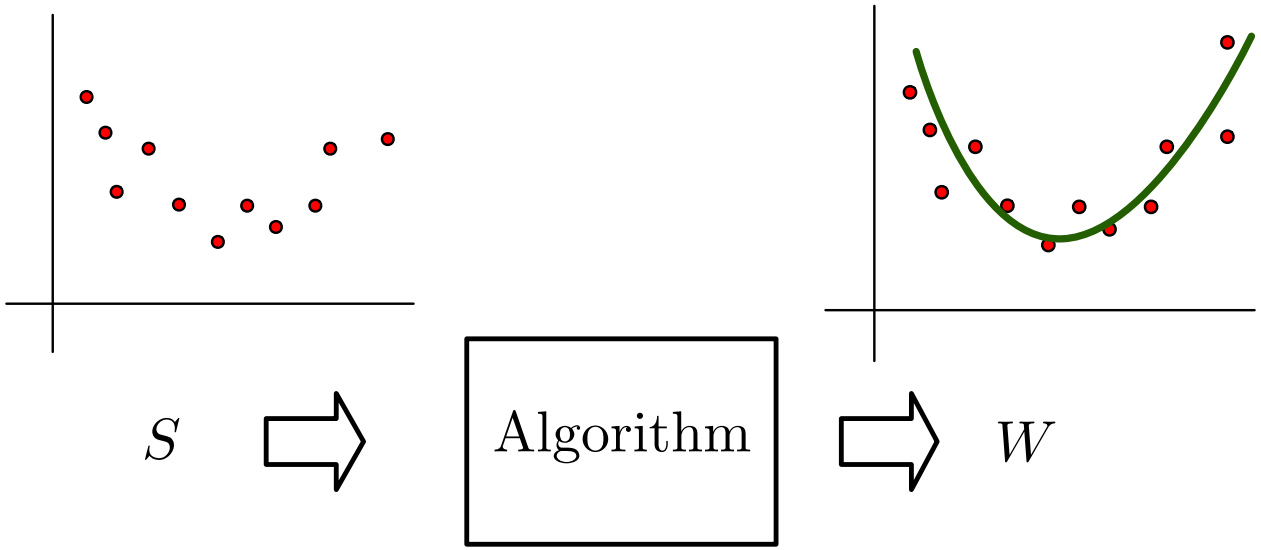
This talk

- Explore information and overfitting connection (Xu & Raginsky, 2017)
- Analyze generalization error in a large and general class of learning algorithms (Pensia, J., Loh, 2018)
- Measuring information via optimal transport theory (Tovar-Lopez, J., 2018)

This talk

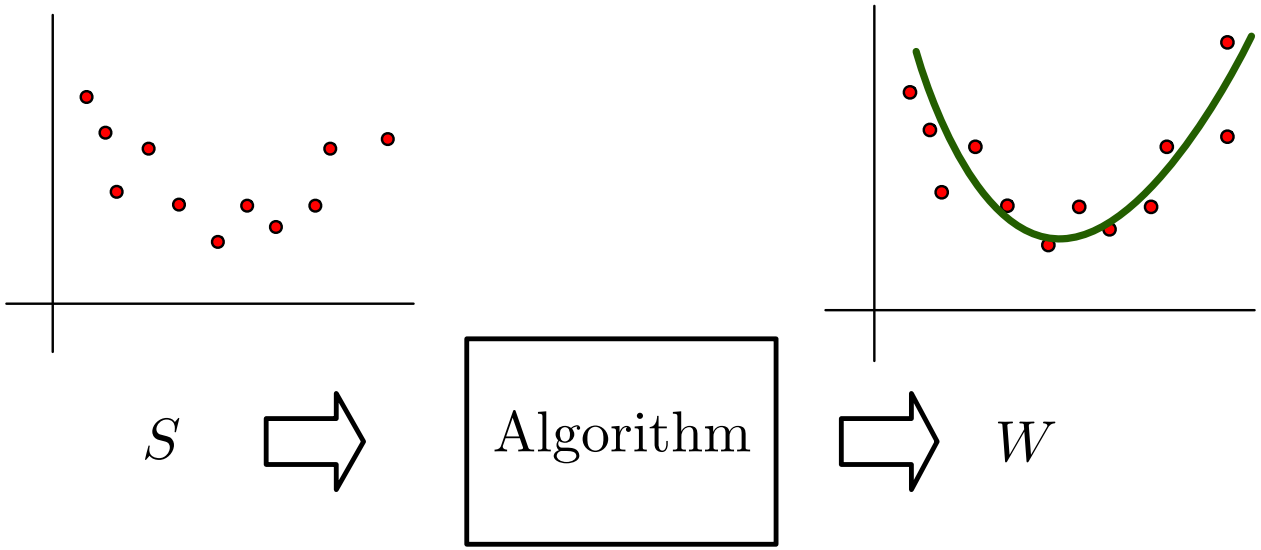
- Explore information and overfitting connection (Xu & Raginsky, 2017)
- Analyze generalization error in a large and general class of learning algorithms (Pensia, J., Loh, 2018)
- Measuring information via optimal transport theory (Tovar-Lopez, J., 2018)
- Speculations and open problems

Learning algorithm as a channel



- Input: Dataset S with N i.i.d. samples $(X_1, X_2, \ldots, X_n) \sim \mu^{\otimes n}$
- Output: W

Learning algorithm as a channel



- Input: Dataset S with N i.i.d. samples $(X_1, X_2, \ldots, X_n) \sim \mu^{\otimes n}$
- Output: W
- Algorithm equivalent to designing $\mathbb{P}_{W|S}$. Very different from channel coding!

Goal of $\mathbb{P}_{W|S}$

- Loss function: $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$

Goal of $\mathbb{P}_{W|S}$

- Loss function: $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$
- Best choice is $w^\star$

$$w^\star = \text{argmin}_{w \in \mathcal{W}} \mathbb{E}_{X \sim \mu}[\ell(w, X)]$$

Goal of $\mathbb{P}_{W|S}$

- Loss function: $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$
- Best choice is $w^\star$

$$w^\star = \text{argmin}_{w \in \mathcal{W}} \mathbb{E}_{X \sim \mu}[\ell(w, X)]$$

- Can't always get what we want...

Goal of $\mathbb{P}_{W|S}$

- Loss function: $\ell : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$
- Best choice is $w^\star$

$$w^\star = \text{argmin}_{w \in \mathcal{W}} \mathbb{E}_{X \sim \mu}[\ell(w, X)]$$

- Can't always get what we want...
- Minimize empirical loss instead

$$\ell_N(w, S) = \frac{1}{N} \sum_{i=1}^{N} \ell(w, X_i)$$

Generalization error

- Define expected loss $= \mathbb{E}_{\substack{X \sim \mu \\ \mathbb{P}_{W|S}\mathbb{P}_S}} \ell(W, X)$ (test error)

Generalization error

- Define expected loss $= \mathbb{E}_{\substack{X \sim \mu \\ \mathbb{P}_{W|S}\mathbb{P}_S}} \ell(W, X)$ (test error)
- Expected empirical loss $= \mathbb{E}_{\mathbb{P}_{WS}} \ell_N(W, S)$ (train error)

Generalization error

- Define expected loss $= \mathbb{E}_{\substack{X\sim\mu \\ \mathbb{P}_{W|S}\mathbb{P}_S}} \ell(W, X)$ (test error)
- Expected empirical loss $= \mathbb{E}_{\mathbb{P}_{WS}} \ell_N(W, S)$ (train error)
- Loss has two parts:

  Expected loss

  $=$ (Expected loss - Expected empirical loss) $+$ Expected empirical loss

  $=$ (test error - train error) $+$ train error

Generalization error

- Define expected loss $= \mathbb{E}_{\substack{X\sim\mu \\ \mathbb{P}_{W|S}\mathbb{P}_S}} \ell(W, X)$ (test error)
- Expected empirical loss $= \mathbb{E}_{\mathbb{P}_{WS}} \ell_N(W, S)$ (train error)
- Loss has two parts:

  Expected loss

  $=$ (Expected loss - Expected empirical loss) $+$ Expected empirical loss

  $=$ (test error - train error) $+$ train error

- Generalization error $=$ test error - train error

$$\text{gen}(\mu, \mathbb{P}_{W|S}) = \mathbb{E}_{\mathbb{P}_S \times \mathbb{P}_W} \ell_N(W, S) - \mathbb{E}_{\mathbb{P}_{WS}} \ell_N(W, S)$$

Generalization error

- Define expected loss $= \mathbb{E}_{\substack{X\sim\mu \\ \mathbb{P}_{W|S}\mathbb{P}_S}} \ell(W, X)$ (test error)
- Expected empirical loss $= \mathbb{E}_{\mathbb{P}_{WS}} \ell_N(W, S)$ (train error)
- Loss has two parts:

  Expected loss

  $=$ (Expected loss - Expected empirical loss) $+$ Expected empirical loss

  $=$ (test error - train error) $+$ train error

- Generalization error $=$ test error - train error

$$\text{gen}(\mu, \mathbb{P}_{W|S}) = \mathbb{E}_{\mathbb{P}_S \times \mathbb{P}_W} \ell_N(W, S) - \mathbb{E}_{\mathbb{P}_{WS}} \ell_N(W, S)$$

- Ideally, we want both small. Often, both are analyzed separately.

Basics of mutual information

- Mutual information $I(X;Y)$ precisely quantifies information between $(X,Y) \sim \mathbb{P}_{XY}$:

$$I(X;Y) = KL(\mathbb{P}_{XY} || \mathbb{P}_X \times \mathbb{P}_Y)$$

Basics of mutual information

- Mutual information $I(X;Y)$ precisely quantifies information between $(X,Y) \sim \mathbb{P}_{XY}$:

$$I(X;Y) = KL(\mathbb{P}_{XY} || \mathbb{P}_X \times \mathbb{P}_Y)$$

- Satisfies two nice properties—

Basics of mutual information

- Mutual information $I(X;Y)$ precisely quantifies information between $(X,Y) \sim \mathbb{P}_{XY}$:

$$I(X;Y) = KL(\mathbb{P}_{XY} || \mathbb{P}_X \times \mathbb{P}_Y)$$

- Satisfies two nice properties—
  - Data processing inequality:

Figure: If $X \to Y \to Z$ then $I(X;Y) \geq I(X;Z)$

Basics of mutual information

- Mutual information $I(X;Y)$ precisely quantifies information between $(X,Y) \sim \mathbb{P}_{XY}$:

$$I(X;Y) = KL(\mathbb{P}_{XY} || \mathbb{P}_X \times \mathbb{P}_Y)$$

- Satisfies two nice properties—
  - Data processing inequality:

Figure: If $X \to Y \to Z$ then $I(X;Y) \geq I(X;Z)$

  - Chain rule:

$$I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X_1)$$

Bounding generalization error using $I(W;S)$

Theorem (Xu & Raginsky (2017))

Assume that $\ell(w, X)$ is R-subgaussian for every $w \in \mathcal{W}$. Then the following bound holds:

$$|gen(\mu, \mathbb{P}_{W|S})| \leq \sqrt{\frac{2R^2}{n} I(S;W)}.$$

How to use it: key insight

How to use it: key insight

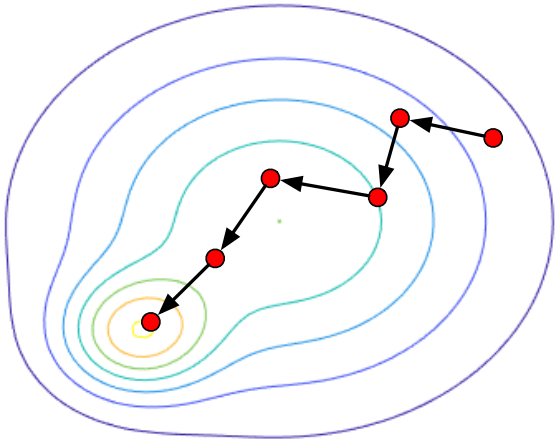


Figure: Update W_t using some update rule to generate W_{t+1}

- Many learning algorithms are iterative
- Generate $W_0, W_1, W_2, \ldots, W_T$, and output $W = f(W_0, \ldots, W_T)$. For example, $W = W_T$ or $W = \frac{1}{T}\sum_i W_i$

How to use it: key insight

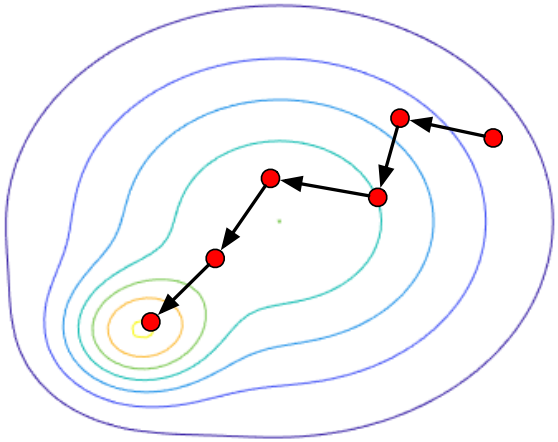


Figure: Update W_t using some update rule to generate W_{t+1}

- Many learning algorithms are iterative
- Generate $W_0, W_1, W_2, \ldots, W_T$, and output $W = f(W_0, \ldots, W_T)$. For example, $W = W_T$ or $W = \frac{1}{T}\sum_i W_i$
- Bound $I(W;S)$ by controlling information at each iteration

Noisy, iterative algorithms

- For $t \geq 1$, sample a minibatch $Z_t \subseteq S$ and compute a direction $F(W_{t-1}, Z_t) \in \mathbb{R}^d$

Noisy, iterative algorithms

- For $t \geq 1$, sample a minibatch $Z_t \subseteq S$ and compute a direction $F(W_{t-1}, Z_t) \in \mathbb{R}^d$
- Move in the direction after scaling by a stepsize η_t

Noisy, iterative algorithms

- For $t \geq 1$, sample a minibatch $Z_t \subseteq S$ and compute a direction $F(W_{t-1}, Z_t) \in \mathbb{R}^d$
- Move in the direction after scaling by a stepsize η_t
- Perturb it by isotropic Gaussian noise $\xi_t \sim N(0, \sigma_t^2 I_d)$

Noisy, iterative algorithms

- For $t \geq 1$, sample a minibatch $Z_t \subseteq S$ and compute a direction $F(W_{t-1}, Z_t) \in \mathbb{R}^d$
- Move in the direction after scaling by a stepsize η_t
- Perturb it by isotropic Gaussian noise $\xi_t \sim N(0, \sigma_t^2 I_d)$
- Overall update equation:

$$W_t = W_{t-1} - \eta_t F(W_{t-1}, Z_t) + \xi_t, \qquad \forall t \geq 1$$

Noisy, iterative algorithms

- For $t \geq 1$, sample a minibatch $Z_t \subseteq S$ and compute a direction $F(W_{t-1}, Z_t) \in \mathbb{R}^d$
- Move in the direction after scaling by a stepsize η_t
- Perturb it by isotropic Gaussian noise $\xi_t \sim N(0, \sigma_t^2 I_d)$
- Overall update equation:

$$W_t = W_{t-1} - \eta_t F(W_{t-1}, Z_t) + \xi_t, \qquad \forall t \geq 1$$

- Run for T steps, output $W = f(W_0, \ldots, W_T)$

Main assumptions

Update equation:

$$W_t = W_{t-1} - \eta_t F(W_{t-1}, Z_t) + \xi_t, \qquad \forall t \geq 1$$

Main assumptions

Update equation:

$$W_t = W_{t-1} - \eta_t F(W_{t-1}, Z_t) + \xi_t, \qquad \forall t \geq 1$$

- Assumption 1: $\ell(w, Z)$ is R-subgaussian

Main assumptions

Update equation:

$$W_t = W_{t-1} - \eta_t F(W_{t-1}, Z_t) + \xi_t, \qquad \forall t \geq 1$$

- Assumption 1: $\ell(w, Z)$ is R-subgaussian
- Assumption 2: Bounded updates; i.e.

$$\sup_{w,z} \|F(w, z)\| \leq L$$

Main assumptions

Update equation:

$$W_t = W_{t-1} - \eta_t F(W_{t-1}, Z_t) + \xi_t, \qquad \forall t \geq 1$$

- Assumption 1: $\ell(w, Z)$ is R-subgaussian
- Assumption 2: Bounded updates; i.e.

$$\sup_{w,z} \|F(w, z)\| \leq L$$

- Assumption 3: Sampling is done without looking at W_t's; i.e.,

$$\mathbb{P}(Z_{t+1} \mid Z^{(t)}, W^{(t)}, S) = \mathbb{P}(Z_{t+1} | Z^{(t)}, S)$$

Main result

Main result

Theorem (Pensia, J., Loh (2018))

The mutual information satisfies the bound

$$I(S;W) \leq \sum_{t=1}^{T} \frac{d}{2} \log\left(1 + \frac{\eta_t^2 L^2}{d\sigma_t^2}\right).$$

Main result

Theorem (Pensia, J., Loh (2018))

The mutual information satisfies the bound

$$I(S;W) \leq \sum_{t=1}^{T} \frac{d}{2} \log\left(1 + \frac{\eta_t^2 L^2}{d\sigma_t^2}\right).$$

- Depends on T — longer you optimize, higher the risk of overfitting

Implications for $\text{gen}(\mu, \mathbb{P}_{W|S})$

Implications for $\text{gen}(\mu, \mathbb{P}_{W|S})$

Corollary (Bound on expectation)

The generalization error of our class of iterative algorithms is bounded by

$$|gen(\mu, P_{W|S})| \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^{T} \frac{\eta_t^2 L^2}{\sigma_t^2}}.$$

Implications for $\text{gen}(\mu, \mathbb{P}_{W|S})$

Corollary (Bound on expectation)

The generalization error of our class of iterative algorithms is bounded by

$$|gen(\mu, P_{W|S})| \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^{T} \frac{\eta_t^2 L^2}{\sigma_t^2}}.$$

Corollary (High-probability bound)

Let $\epsilon = \sum_{t=1}^{T} \frac{d}{2} \log\left(1 + \frac{\eta_t^2 L^2}{d\sigma_t^2}\right)$. *For any* $\alpha > 0$ *and* $0 < \beta \leq 1$, *if* $n > \frac{8R^2}{\alpha^2}\left(\frac{\epsilon}{\beta} + \log(\frac{2}{\beta})\right)$, *we have*

$$\mathbb{P}_{S,W}\left(|L_\mu(W) - L_S(W)| > \alpha\right) \leq \beta, \tag{1}$$

where the probability is with respect to $S \sim \mu^{\otimes n}$ *and* W.

Applications: SGLD

- SGLD iterates are

$$W_{t+1} = W_t - \eta_t \nabla \ell(W_t, Z_t) + \sigma_t Z_t$$

Applications: SGLD

- SGLD iterates are

$$W_{t+1} = W_t - \eta_t \nabla \ell(W_t, Z_t) + \sigma_t Z_t$$

- Common experimental practices for SGLD [Welling & Teh, 2011]:
  1. the noise variance $\sigma_t^2 = \eta_t$,
  2. the algorithm is run for K epochs; i.e., $T = nK$,
  3. for a constant $c > 0$, the stepsizes are $\eta_t = \frac{c}{t}$.

Applications: SGLD

- SGLD iterates are

$$W_{t+1} = W_t - \eta_t \nabla \ell(W_t, Z_t) + \sigma_t Z_t$$

- Common experimental practices for SGLD [Welling & Teh, 2011]:
  1. the noise variance $\sigma_t^2 = \eta_t$,
  2. the algorithm is run for K epochs; i.e., $T = nK$,
  3. for a constant $c > 0$, the stepsizes are $\eta_t = \frac{c}{t}$.
- Expectation bounds: Using $\sum_{t=1}^{T} \frac{1}{t} \leq \log(T) + 1$

$$|\text{gen}(\mu, \mathbb{P}_{W|S})| \leq \frac{RL}{\sqrt{n}} \sqrt{\sum_{t=1}^{T} \eta_t} \leq \frac{RL}{\sqrt{n}} \sqrt{c \log T + c}$$

Applications: SGLD

- SGLD iterates are

$$W_{t+1} = W_t - \eta_t \nabla \ell(W_t, Z_t) + \sigma_t Z_t$$

- Common experimental practices for SGLD [Welling & Teh, 2011]:
  1. the noise variance $\sigma_t^2 = \eta_t$,
  2. the algorithm is run for K epochs; i.e., $T = nK$,
  3. for a constant $c > 0$, the stepsizes are $\eta_t = \frac{c}{t}$.
- Expectation bounds: Using $\sum_{t=1}^{T} \frac{1}{t} \leq \log(T) + 1$

$$|\mathrm{gen}(\mu, \mathbb{P}_{W|S})| \leq \frac{RL}{\sqrt{n}} \sqrt{\sum_{t=1}^{T} \eta_t} \leq \frac{RL}{\sqrt{n}} \sqrt{c \log T + c}$$

- Best known bounds by Mou et al. (2017) are $O(1/n)$—but our bounds more general

Other applications

- Perturbed SGD (Ge et al. 2015): Similar to SGLD, but different noise distribution:

$$W_t = W_{t-1} - \eta \left(\nabla_w \ell(W_{t-1}, Z_t) + \xi_t \right),$$

where $\xi_t \sim \text{Unif}(\mathcal{B}_d)$ (unit ball in $\mathbb{R}^d$)

Other applications

- Perturbed SGD (Ge et al. 2015): Similar to SGLD, but different noise distribution:

$$W_t = W_{t-1} - \eta \left(\nabla_w \ell(W_{t-1}, Z_t) + \xi_t \right),$$

where $\xi_t \sim \text{Unif}(\mathcal{B}_d)$ (unit ball in $\mathbb{R}^d$)

- A modified version of stochastic gradient Hamiltonian Monte-Carlo, Chen et al. (2014):

$$\begin{aligned} V_t &= \gamma_t V_{t-1} + \eta_t \nabla_w \ell(W_{t-1}, Z_t) + \xi'_t, \\ W_t &= W_{t-1} - \gamma_t V_{t-1} - \eta_t \nabla_w \ell(W_{t-1}, Z_t) + \xi''_t, \end{aligned}$$

Other applications

- Perturbed SGD (Ge et al. 2015): Similar to SGLD, but different noise distribution:

$$W_t = W_{t-1} - \eta \left(\nabla_w \ell(W_{t-1}, Z_t) + \xi_t \right),$$

  where $\xi_t \sim \text{Unif}(\mathcal{B}_d)$ (unit ball in $\mathbb{R}^d$)
- A modified version of stochastic gradient Hamiltonian Monte-Carlo, Chen et al. (2014):

$$\begin{aligned} V_t &= \gamma_t V_{t-1} + \eta_t \nabla_w \ell(W_{t-1}, Z_t) + \xi_t', \\ W_t &= W_{t-1} - \gamma_t V_{t-1} - \eta_t \nabla_w \ell(W_{t-1}, Z_t) + \xi_t'', \end{aligned}$$

- Same bound also holds for "noisy" Nesterov's accelerated gradient descent method (1983)

Proof sketch

Lots of Markov chains!

Proof sketch

Lots of Markov chains!

- $I(W;S) \leq I(W_0^T; Z_1^T)$ because

$$S \to Z_1^T \to W_0^T \to W$$

Figure: Data processing inequality

- Iterative structure means

$$W_0 \to Z_1 W_1 \to Z_2 W_2 \to Z_3 W_3 \cdots \to W_T$$

Proof sketch

Lots of Markov chains!

- $I(W;S) \leq I(W_0^T; Z_1^T)$ because

$$S \to Z_1^T \to W_0^T \to W$$

Figure: Data processing inequality

- Iterative structure means

$$W_0 \to Z_1 W_1 \to Z_2 W_2 \to Z_3 W_3 \cdots \to W_T$$

- Use Markovity with chain rule to get

$$I(Z_1^T; W_0^T) = \sum_{t=1}^{T} I(Z_t; W_t | W_{t-1})$$

Proof sketch

Lots of Markov chains!

- $I(W;S) \leq I(W_0^T; Z_1^T)$ because

$$S \to Z_1^T \to W_0^T \to W$$

Figure: Data processing inequality

- Iterative structure means

$$W_0 \to Z_1 W_1 \to Z_2 W_2 \to Z_3 W_3 \cdots \to W_T$$

- Use Markovity with chain rule to get

$$I(Z_1^T; W_0^T) = \sum_{t=1}^{T} I(Z_t; W_t | W_{t-1})$$

- Bottomline: Bound "one step" information between W_t and Z_t

Proof sketch

- Recall

$$W_t = W_{t-1} - \eta_t F(W_{t-1}, Z_t) + \xi_t$$

Proof sketch

- Recall

$$W_t = W_{t-1} - \eta_t F(W_{t-1}, Z_t) + \xi_t$$

- Using the entropy form of mutual information,

$$I(W_t; Z_t | W_{t-1}) = \underbrace{h(W_t|W_{t-1})}_{Variance(W_t|w_{t-1}) \leq \eta_t^2 L^2 + \sigma_t^2} - \underbrace{h(W_t|W_{t-1}, Z_t)}_{=h(\xi_t)}$$

Proof sketch

- Recall

$$W_t = W_{t-1} - \eta_t F(W_{t-1}, Z_t) + \xi_t$$

- Using the entropy form of mutual information,

$$I(W_t; Z_t | W_{t-1}) = \underbrace{h(W_t|W_{t-1})}_{\text{Variance}(W_t|w_{t-1}) \leq \eta_t^2 L^2 + \sigma_t^2} - \underbrace{h(W_t|W_{t-1}, Z_t)}_{= h(\xi_t)}$$

- Gaussian distribution maximizes entropy for fixed variance, giving

$$I(W_t; Z_t | W_{t-1}) \leq \frac{d}{2} \log \left(1 + \frac{\eta_t^2 L^2}{d\sigma_t^2} \right)$$

What's wrong with mutual information

What's wrong with mutual information

- Mutual information is great, but ...

What's wrong with mutual information

- Mutual information is great, but ...
- ...many cases when mutual information $I(W; S)$ shoots to infinity

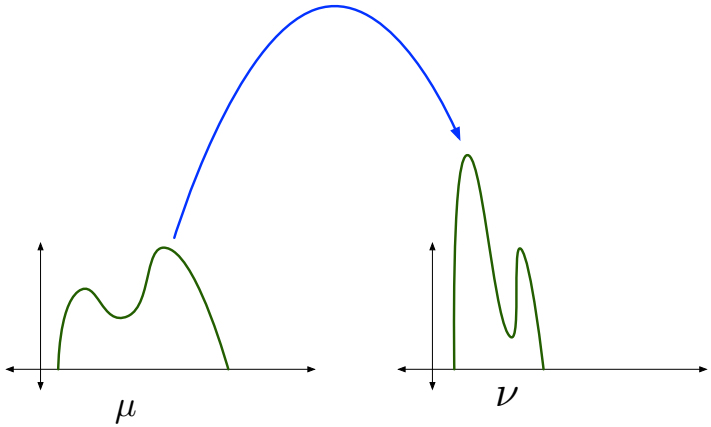
What's wrong with mutual information

- Mutual information is great, but ...
- ...many cases when mutual information $I(W;S)$ shoots to infinity
- Cannot use bounds for stochastic gradient descent (SGD) :(

What's wrong with mutual information

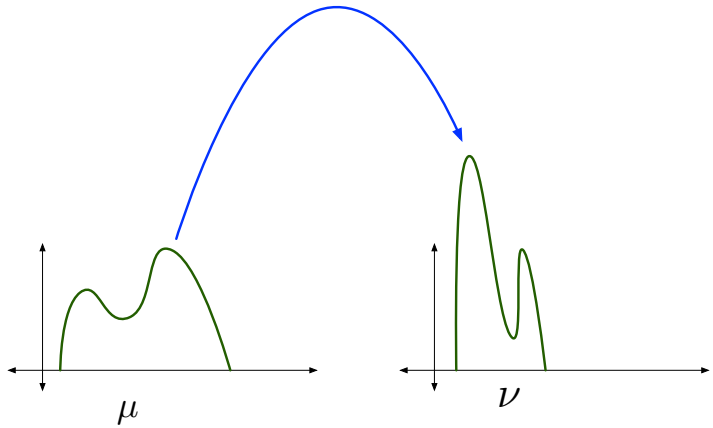
- Mutual information is great, but ...
- ...many cases when mutual information $I(W;S)$ shoots to infinity
- Cannot use bounds for stochastic gradient descent (SGD) :(
- "Noisy" algorithms are *essential* for using mutual information based bounds

Wasserstein metric



[1] Topics in Optimal Transportation by Cedric Villani

Wasserstein metric



- Wasserstein distance given by

$$W_p(\mu,\nu) = \left(\inf_{\mathbb{P}_{XY} \in \Pi(\mu,\nu)} \mathbb{E}\|X - Y\|^p \right)^{1/p}$$

where $\Pi(\mu,\nu)$ is the set of coupling such that marginals are μ and ν

- Lots of fascinating theory[1] for W_p

[1] Topics in Optimal Transportation by Cedric Villani

Main result: Wasserstein bounds on $\text{gen}(\mu, \mathbb{P}_{W|S})$

- Assumption: $\ell(w, x)$ is Lipschitz in x for each fixed w; i.e.

$$|\ell(w, x_1) - \ell(w, x_2)| \leq L\|x_1 - x_2\|_p$$

Main result: Wasserstein bounds on $\text{gen}(\mu, \mathbb{P}_{W|S})$

- Assumption: $\ell(w, x)$ is Lipschitz in x for each fixed w; i.e.

$$|\ell(w, x_1) - \ell(w, x_2)| \leq L\|x_1 - x_2\|_p$$

Theorem (Tovar-Lopez & J., (2018))

If $\ell(w, \cdot)$ is L-Lipschitz in $\|\cdot\|_p$, generalization error satisfies the following bound:

$$gen(\mu, \mathbb{P}_{W|S}) \leq \frac{L}{n^{\frac{1}{p}}} \left(\int_W W_p^p(\mathbb{P}_S, \mathbb{P}_{S|w}) d\mathbb{P}_W(w) \right)^{\frac{1}{p}}$$

Main result: Wasserstein bounds on $\text{gen}(\mu, \mathbb{P}_{W|S})$

- Assumption: $\ell(w, x)$ is Lipschitz in x for each fixed w; i.e.

$$|\ell(w, x_1) - \ell(w, x_2)| \leq L\|x_1 - x_2\|_p$$

Theorem (Tovar-Lopez & J., (2018))

If $\ell(w, \cdot)$ is L-Lipschitz in $\|\cdot\|_p$, generalization error satisfies the following bound:

$$gen(\mu, \mathbb{P}_{W|S}) \leq \frac{L}{n^{\frac{1}{p}}} \left(\int_W W_p^p(\mathbb{P}_S, \mathbb{P}_{S|w}) d\mathbb{P}_W(w) \right)^{\frac{1}{p}}$$

- Measure average separation of $\mathbb{P}_{S|W}$ from $\mathbb{P}_S$ (looks like a p-th moment in the space of distributions)

Comparison of two bounds

Comparison of two bounds

- In general, not comparable

Comparison of two bounds

- In general, not comparable
- If μ satisfies a $T_p(c)$-transportation inequality, for $p \in [1,2]$, then can directly compare:

Theorem (Tovar-Lopez & J., (2018))

Assume that μ satisfies $T_p(c)$ for some $p \in [1,2]$ and some $c > 0$. Then

$$\frac{L}{n^{\frac{1}{p}}} \left(\int_W W_p^p(\mathbb{P}_S, \mathbb{P}_{S|w}) d\mathbb{P}_W(w) \right)^{\frac{1}{p}} \leq L\sqrt{\frac{2c}{n} I(S;W)}$$

Comparison of two bounds

- In general, not comparable
- If μ satisfies a $T_p(c)$-transportation inequality, for $p \in [1,2]$, then can directly compare:

Theorem (Tovar-Lopez & J., (2018))

Assume that μ satisfies $T_p(c)$ for some $p \in [1,2]$ and some $c > 0$. Then

$$\frac{L}{n^{\frac{1}{p}}} \left(\int_W W_p^p(\mathbb{P}_S, \mathbb{P}_{S|w}) d\mathbb{P}_W(w) \right)^{\frac{1}{p}} \leq L\sqrt{\frac{2c}{n} I(S;W)}$$

- In particular, for Gaussian data, Wasserstein bound strictly stronger

Proof sketch

- Recall generalization error expression:

$$\text{gen}(\mu, \mathbb{P}_{W|S}) = |\mathbb{E}\ell_N(\bar{S}, \bar{W}) - \mathbb{E}\ell_N(S, W)|,$$

where $(\bar{S}, \bar{W}) \sim \mathbb{P}_S \times \mathbb{P}_W$ and $(S, W) \sim \mathbb{P}_{WS}$.

Proof sketch

- Recall generalization error expression:

$$\text{gen}(\mu, \mathbb{P}_{W|S}) = |\mathbb{E}\ell_N(\bar{S}, \bar{W}) - \mathbb{E}\ell_N(S, W)|,$$

  where $(\bar{S}, \bar{W}) \sim \mathbb{P}_S \times \mathbb{P}_W$ and $(S, W) \sim \mathbb{P}_{WS}$.
- Key insight: Any coupling of $(\bar{S}, \bar{W}, S, W)$ that has the "correct" marginals on (S, W) and $(\bar{S}, \bar{W})$ leads to the same expected value above

Proof sketch

- Recall generalization error expression:

$$\text{gen}(\mu, \mathbb{P}_{W|S}) = |\mathbb{E}\ell_N(\bar{S}, \bar{W}) - \mathbb{E}\ell_N(S, W)|,$$

  where $(\bar{S}, \bar{W}) \sim \mathbb{P}_S \times \mathbb{P}_W$ and $(S, W) \sim \mathbb{P}_{WS}$.
- Key insight: Any coupling of $(\bar{S}, \bar{W}, S, W)$ that has the "correct" marginals on (S, W) and $(\bar{S}, \bar{W})$ leads to the same expected value above
- Just pick a nice coupling!

Speculations and open problems

Speculations and open problems

- Stability: How much does W change with S changes a little?
- Property of the forward channel $\mathbb{P}_{W|S}$

Speculations and open problems

- Stability: How much does W change with S changes a little?
- Property of the forward channel $\mathbb{P}_{W|S}$
- Generalization upper bound: How much does S change when W changes a little?
- Property of the backward channel $\mathbb{P}_{S|W}$

Speculations and open problems

- Stability: How much does W change with S changes a little?
- Property of the forward channel $\mathbb{P}_{W|S}$
- Generalization upper bound: How much does S change when W changes a little?
- Property of the backward channel $\mathbb{P}_{S|W}$
- Pre-process data to deliberately make backward channel noisy (data augmentation, smoothing, etc.)

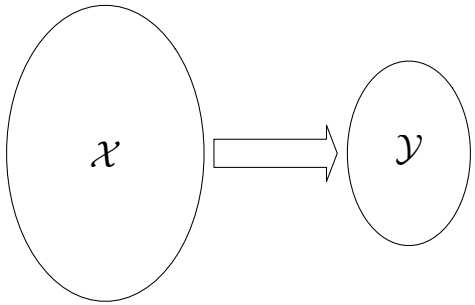
Speculations and open problems

Relation to rate distortion theory?

Speculations and open problems

Relation to rate distortion theory?

- Branch of information theory dealing with lossy data compression

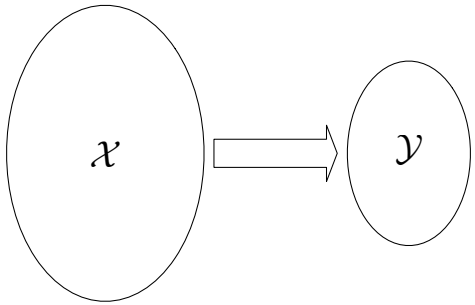


$$\min_{\mathbb{P}_{Y|X}} \mathbb{E}d(X,Y) \text{ subject to } I(X;Y) \leq R$$

Speculations and open problems

Relation to rate distortion theory?

- Branch of information theory dealing with lossy data compression



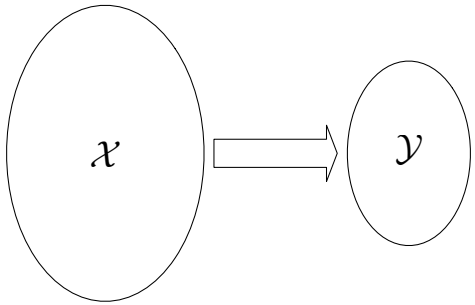
$$\min_{\mathbb{P}_{Y|X}} \mathbb{E}d(X,Y) \text{ subject to } I(X;Y) \leq R$$

- Minimize distortion given by $\ell_N(W,S)$ subject to mutual information constraint $I(W;S) \leq \epsilon$

Speculations and open problems

Relation to rate distortion theory?

- Branch of information theory dealing with lossy data compression



$$\min_{\mathbb{P}_{Y|X}} \mathbb{E}d(X,Y) \text{ subject to } I(X;Y) \leq R$$

- Minimize distortion given by $\ell_N(W,S)$ subject to mutual information constraint $I(W;S) \leq \epsilon$
- Essentially same problem, but more connections need to be developed

References

- *Generalization error for noisy, iterative algorithms*. Ankit Pensia, Varun Jog, Po-Ling Loh. ISIT 2018.
- *Wasserstein distance based bounds on generalization error*. Adrian Tovar Lopez, Varun Jog. ITW 2018.

Thank you!