Studying generalization in deep learning via PAC-Bayes

Gintarė Karolina Džiugaitė — Element AI

Joint work with

Daniel M. Roy — University of Toronto; Vector Institute
Kyle Hsu — University of Toronto; Vector Institute
Waseem Gharbieh — Element AI
Motivation: SGD is widely used in practice yet its generalization properties are poorly understood.
Motivation: SGD is widely used in practice yet its generalization properties are poorly understood.

Approach: PAC-Bayes:
Motivation: SGD is widely used in practice yet its generalization properties are poorly understood.

Approach: PAC-Bayes:

- PAC-Bayes yields bounds on average risk for weights $\sim Q = Q(S)$;
Motivation: SGD is widely used in practice yet its generalization properties are poorly understood.

Approach: PAC-Bayes:

- PAC-Bayes yields bounds on average risk for weights $\sim Q = Q(S)$;
- in order to study SGD, need “posterior” $Q$ concentrated near SGD solution;
Motivation: SGD is widely used in practice yet its generalization properties are poorly understood.

Approach: PAC-Bayes:

- PAC-Bayes yields bounds on average risk for weights $\sim Q = Q(S)$;
- in order to study SGD, need “posterior” $Q$ concentrated near SGD solution;
- generalization error bound is then determined by $\text{KL}(Q||P)$, where “prior” $P$ is fixed.
**Motivation**: SGD is widely used in practice yet its generalization properties are poorly understood.

**Approach**: PAC-Bayes:

- PAC-Bayes yields bounds on average risk for weights $\sim Q = Q(S)$;
- in order to study SGD, need “posterior” $Q$ concentrated near SGD solution;
- generalization error bound is then determined by $\text{KL}(Q||P)$, where “prior” $P$ is fixed.
- Empirically, existing bounds are numerically vacuous ($> 1$) for numerous reasons: almost all applications suffer from large KL divergence on the account of bad choice of $P$. 

IAS Theory of Deep Learning Workshop 2019
**Motivation:** SGD is widely used in practice yet its generalization properties are poorly understood.

**Approach:** PAC-Bayes:

- PAC-Bayes yields bounds on average risk for weights $\sim Q = Q(S)$;
- in order to study SGD, need “posterior” $Q$ concentrated near SGD solution;
- generalization error bound is then determined by $\text{KL}(Q||P)$, where “prior” $P$ is fixed.
- Empirically, existing bounds are numerically vacuous ($> 1$) for numerous reasons: almost all applications suffer from large KL divergence on the account of bad choice of $P$.
- I’ll focus on the role of the prior $P$. 
Review the PAC-Bayes framework for generalization bounds.

Introduce three principles for studying generalization using PAC-Bayes framework.

Describe their application to computing risk bounds on $Q_{SGD}$.

Show how same ideas can be applied to self-bounded learning.
Review the PAC-Bayes framework for generalization bounds.
Outline

- Review the PAC-Bayes framework for generalization bounds.
- Introduce three principles for studying generalization using PAC-Bayes framework.
Review the PAC-Bayes framework for generalization bounds.

Introduce three principles for studying generalization using PAC-Bayes framework.

Describe their application to computing risk bounds on $Q$ concentrated near $w_{SGD}$. 

Show how same ideas can be applied to self-bounded learning.
Outline

- Review the PAC-Bayes framework for generalization bounds.
- Introduce three principles for studying generalization using PAC-Bayes framework.
- Describe their application to computing risk bounds on $Q$ concentrated near $w_{\text{SGD}}$.
- Show how same ideas can be applied to self-bounded learning.
PAC-Bayes yields risk bounds for Gibbs classifiers

Let $\mathcal{H}$ be weight space (which determine classifiers).
Let $\ell : \mathcal{H} \times Z \to [0, 1]$ be our loss function.

Risk and empirical risk

For $h \in \mathcal{H}$,

$$L_D(h) = \mathbb{E}_{z \sim D}[\ell(h, z)] \quad \text{risk}$$

$$L_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h, z_i) \quad \text{empirical risk}$$

Gibbs classifier

A Gibbs classifier is a probability distribution on $\mathcal{H}$.

The risk of a Gibbs classifier $Q$ is defined to be the average risk under $w \sim Q$, i.e.,

$$L_D(Q) = \mathbb{E}_{h \sim Q}[L_D(h)] = \mathbb{E}_{z \sim D} \mathbb{E}_{h \sim Q}[\ell(h, z)].$$
PAC-Bayes generalization bounds

**Theorem** (PAC-Bayes; Catoni 2007)
McAllester 1999, Shawe-Taylor and Williamson 1997
Assume $L_S(\cdot) \in [0, 1]$. 

1. Nature chooses a data distribution $D$.
2. We choose a distribution $P$ on weights (the “prior”).
3. Nature gives us a data set $S \sim D^m$.
4. Then, with probability at least $(1 - \delta)$,
PAC-Bayes generalization bounds

**Theorem** (PAC-Bayes; Catoni 2007)

McAllester 1999, Shawe-Taylor and Williamson 1997

Assume $L_S(\cdot) \in [0, 1]$.

1. Nature chooses a data distribution $\mathcal{D}$.

\[ \mathbb{R}^d \]
Theorem (PAC-Bayes; Catoni 2007)
McAllester 1999, Shawe-Taylor and Williamson 1997
Assume \( L_S(\cdot) \in [0, 1] \).

1. Nature chooses a data distribution \( D \).
2. We choose a distribution \( P \) on weights (the “prior”).
PAC-Bayes generalization bounds

**Theorem** (PAC-Bayes; Catoni 2007)

McAllester 1999, Shawe-Taylor and Williamson 1997

Assume \( L_S(\cdot) \in [0, 1] \).

1. Nature chooses a data distribution \( D \).
2. We choose a distribution \( P \) on weights (the “prior”).
3. Nature gives us a data set \( S \sim D^m \).
   Now we know the empirical risk surface \( L_S(\cdot) \).
Theorem (PAC-Bayes; Catoni 2007)

McAllester 1999, Shawe-Taylor and Williamson 1997

Assume $L_S(\cdot) \in [0, 1]$.

1. Nature chooses a data distribution $\mathcal{D}$.
2. We choose a distribution $P$ on weights (the “prior”).
3. Nature gives us a data set $S \sim \mathcal{D}^m$.
   Now we know the empirical risk surface $L_S(\cdot)$. 
Theorem (PAC-Bayes; Catoni 2007)

McAllester 1999, Shawe-Taylor and Williamson 1997

Assume $L_S(\cdot) \in [0, 1]$.

1. Nature chooses a data distribution $\mathcal{D}$.
2. We choose a distribution $P$ on weights (the “prior”).
3. Nature gives us a data set $S \sim \mathcal{D}^m$.
   Now we know the empirical risk surface $L_S(\cdot)$.
4. Then, with probability at least $(1 - \delta)$,

   $$\forall Q, \quad L_D(Q) \leq L_S(Q) + \sqrt{\frac{\text{KL}(Q||P) + \ln m/\delta}{2m}}.$$
Theorem (PAC-Bayes; Catoni 2007)

Assume $L_S(\cdot) \in [0, 1]$.

1. Nature chooses a data distribution $\mathcal{D}$.
2. We choose a distribution $P$ on weights (the “prior”).
3. Nature gives us a data set $S \sim \mathcal{D}^m$.
   Now we know the empirical risk surface $L_S(\cdot)$.
4. Then, with probability at least $(1 - \delta)$,

$$\forall Q, \ L_D(Q) \leq L_S(Q) + \frac{\sqrt{\text{KL}(Q||P) + \ln m/\delta}}{2m}.$$
Theorem (PAC-Bayes; Catoni 2007)

McAllester 1999, Shawe-Taylor and Williamson 1997

Assume $L_S(\cdot) \in [0, 1]$.

1. Nature chooses a data distribution $D$.
2. We choose a distribution $P$ on weights (the “prior”).
3. Nature gives us a data set $S \sim D^m$.
   
   Now we know the empirical risk surface $L_S(\cdot)$.
4. Then, with probability at least $(1 - \delta)$,

$$\forall Q, \ L_D(Q) \leq L_S(Q) + \sqrt{\frac{\text{KL}(Q||P) + \ln m/\delta}{2m}}.$$
PAC-Bayes generalization bounds

**Theorem** (PAC-Bayes; Catoni 2007)

McAllester 1999, Shawe-Taylor and Williamson 1997

Assume $L_S(\cdot) \in [0, 1]$.

1. Nature chooses a data distribution $\mathcal{D}$.
2. We choose a distribution $P$ on weights (the “prior”).
3. Nature gives us a data set $S \sim \mathcal{D}^m$.
   
   Now we know the empirical risk surface $L_S(\cdot)$.
4. Then, with probability at least $(1 - \delta)$,

\[
\forall Q, \quad L_D(Q) \leq L_S(Q) + \sqrt{\frac{\text{KL}(Q||P) + \ln m/\delta}{2m}}.
\]

\[
\forall Q, \quad 2(L_D(Q) - L_S(Q))^2 \leq \frac{\text{KL}(Q||P) + \ln m/\delta}{m}.
\]
**Theorem** (PAC-Bayes; Catoni 2007)

McAllester 1999, Shawe-Taylor and Williamson 1997

Assume $L_S(\cdot) \in [0, 1]$.

1. Nature chooses a data distribution $\mathcal{D}$.
2. We choose a distribution $P$ on weights (the “prior”).
3. Nature gives us a data set $S \sim \mathcal{D}^m$.
   Now we know the empirical risk surface $L_S(\cdot)$.
4. Then, with probability at least $(1 - \delta)$,

\[
\forall Q, \quad L_D(Q) \leq L_S(Q) + \sqrt{\frac{\text{KL}(Q||P) + \ln m/\delta}{2m}}.
\]

\[
\forall Q, \quad \Delta(L_S(Q), L_D(Q)) \leq \frac{\text{KL}(Q||P) + \ln I^\Delta(m)/\delta}{m}.
\]
Growing literature on techniques to construct PAC-Bayes bounds on deterministic classifiers.
Growing literature on techniques to construct PAC-Bayes bounds on deterministic classifiers.

- **Exploiting margin to derandomize**
  - Herbrich and Graepel (2001)
  - Neyshabur et al. (2019)
  - Nagarajan and Kolter (2019)
PAC-Bayes bounds on deterministic classifiers

Growing literature on techniques to construct PAC-Bayes bounds on deterministic classifiers.

- **Exploiting margin to derandomize**
Growing literature on techniques to construct PAC-Bayes bounds on deterministic classifiers.

- **Exploiting margin to derandomize**

**Theorem (Neyshabur et al. 2019).** Fix margin $\gamma > 0$ and confidence $\delta > 0$. For each $h \in \mathcal{H}$, let $Q(h)$ be a distribution on $\mathcal{H}$ satisfying, with probability $\geq \frac{1}{2}$ over $h' \sim Q(H)$,

$$\sup_z \| f_h(z) - f_{h'}(z) \|_\infty \leq \frac{\gamma}{4}. $$

Then, with probability at least $(1 - \delta)$,

$$\forall h \in \mathcal{H}, L_D(h) \leq L_\gamma(h) + 4\sqrt{\frac{\text{KL}(Q(h)\|P) + \ln \frac{6m}{\delta}}{m + 1}}$$
Growing literature on techniques to construct PAC-Bayes bounds on deterministic classifiers.

- Exploiting margin to derandomize
PAC-Bayes bounds on deterministic classifiers

Growing literature on techniques to construct PAC-Bayes bounds on deterministic classifiers.

- **Exploiting margin to derandomize**

- **Disintegrated versions of PAC-Bayes**
  Catoni (2007)

- ...

- **PAC-Bayes + Generic Chaining**
  Miyaguchi (2019)
Recap: Towards a nonvacuous bound on SGD

\[ \forall Q, \Delta \left( L_S(Q), L_D(Q) \right) \leq \frac{\text{KL}(Q||P) + \ln \mathcal{I}^\Delta(m)/\delta}{m}. \]

Consider a PAC-Bayes + margin approach to bounding SGD risk:
Recap: Towards a nonvacuous bound on SGD

\[ \forall Q, \Delta(L_S(Q), L_D(Q)) \leq \frac{\text{KL}(Q||P) + \ln \mathcal{I}^\Delta(m)/\delta}{m}. \]

Consider a PAC-Bayes + margin approach to bounding SGD risk:

▶ **In order to derandomize...**

Need posterior \( Q \) tightly concentrated around weights \( w_{\text{SGD}} \) learned by SGD.
Recap: Towards a nonvacuous bound on SGD

\[ \forall Q, \Delta(L_S(Q), L_D(Q)) \leq \frac{\text{KL}(Q||P) + \ln \mathcal{I}^\Delta(m)/\delta}{m}. \]

Consider a PAC-Bayes + margin approach to bounding SGD risk:

- **In order to derandomize...**
  Need posterior Q tightly concentrated around weights \( w_{\text{SGD}} \) learned by SGD.

- **In order to control KL complexity term...**
  Need prior P to have sufficient mass near SGD solution.
Recap: Towards a nonvacuous bound on SGD

\[ \forall Q, \Delta\left(L_S(Q), L_D(Q)\right) \leq \frac{KL(Q\|P) + \ln \mathcal{I}(\Delta(m)/\delta)}{m}. \]

Consider a PAC-Bayes + margin approach to bounding SGD risk:

- **In order to derandomize...**
  Need posterior \( Q \) tightly concentrated around weights \( w_{SGD} \) learned by SGD.

- **In order to control KL complexity term...**
  Need prior \( P \) to have sufficient mass near SGD solution.

Hard to achieve both at the same time without knowing the training data \( S \) or at least the data distribution \( \mathcal{D} \).
∀Q, \( \Delta \left( L_S(Q), L_D(Q) \right) \leq \frac{\text{KL}(Q||P) + \ln I^\Delta(m)/\delta}{m} \).

Consider a PAC-Bayes + margin approach to bounding SGD risk:

- **In order to derandomize...**
  Need posterior \( Q \) tightly concentrated around weights \( w_{\text{SGD}} \) learned by SGD.

- **In order to control KL complexity term...**
  Need prior \( P \) to have sufficient mass near SGD solution.

**Hard to achieve both at the same time without knowing the training data \( S \) or at least the data distribution \( \mathcal{D} \).** In fact, the prior can depend on the data distribution!
Recap: Towards a nonvacuous bound on SGD

\[ \forall Q, \quad \Delta(L_S(Q), L_D(Q)) \leq \frac{\text{KL}(Q\|P) + \ln I^\Delta(m)/\delta}{m}. \]

Consider a PAC-Bayes + margin approach to bounding SGD risk:

- **In order to derandomize...**
  Need posterior \( Q \) tightly concentrated around weights \( w_{\text{SGD}} \) learned by SGD.

- **In order to control KL complexity term...**
  Need prior \( P \) to have sufficient mass near SGD solution.

Hard to achieve both at the same time without knowing the training data \( S \) or at least the data distribution \( \mathcal{D} \). In fact, the prior can depend on the data distribution!

**Theorem (Catoni 2007; Langford).** "Optimal" prior is \( P^* = \mathbb{E}_{S \sim \mathcal{D}^m}[Q(S)] \).
Recap: Towards a nonvacuous bound on SGD

\[ \forall Q, \Delta \left( L_S(Q), L_D(Q) \right) \leq \frac{\text{KL}(Q \| P) + \ln I^\Delta(m)/\delta}{m}. \]

Consider a PAC-Bayes + margin approach to bounding SGD risk:

- **In order to derandomize...**
  Need posterior \( Q \) tightly concentrated around weights \( w_{\text{SGD}} \) learned by SGD.

- **In order to control KL complexity term...**
  Need prior \( P \) to have sufficient mass near SGD solution.

Hard to achieve both at the same time without knowing the training data \( S \) or at least the data distribution \( \mathcal{D} \). In fact, the prior *can* depend on the data distribution!

**Theorem (Catoni 2007; Langford).** "Optimal" prior is \( P^* = \mathbb{E}_{S \sim \mathcal{D}^m}[Q(S)] \).

\[ \mathbb{E}_{S \sim \mathcal{D}^m}[\text{KL}(Q(S) \| P^*)] = I(S; W) \text{ where } W \mid S \sim Q(S). \]
Can we exploit optimal priors?

Optimal prior $P^* = \mathbb{E}_{S' \sim D}[Q(S')]$ depends on $D$.

**Fundamental tension**

1. PAC-Bayes prior $P$ can depend on data distribution $D$ but cannot depend on the data $S$.
2. Our only handle on the unknown distribution $D$ is the sample $S$.

▶ Distribution-dependent priors + KL bounds
Catoni (2004; 2007); Lever et al. (2010); Rivasplata et al. (2019)

▶ Data-dependent priors
Use a subset of data to learn prior
Use remainder of data for bound (Ambroladze et al. 2007; Parrado-Hernández et al. 2012)

▶ Use all the data + differential privacy
(D. and Roy 2018a)
Can we exploit optimal priors?

Optimal prior $P^* = \mathbb{E}_{S' \sim D^m}[Q(S')]$ depends on $D$. 

Fundamental tension
1. PAC-Bayes prior $P$ can depend on data distribution $D$ but cannot depend on the data $S$;
2. Our only handle on the unknown distribution $D$ is the sample $S$.

▶ Distribution-dependent priors + KL bounds
Catoni (2004; 2007); Lever et al. (2010); Rivasplata et al. (2019)

▶ Data-dependent priors
▶ Use a subset of data to learn prior
Use remainder of data for bound (Ambroladze et al. 2007; Parrado-Hernández et al. 2012)

▶ Use all the data + differential privacy
(D. and Roy 2018a)
Can we exploit optimal priors?

Optimal prior $P^* = \mathbb{E}_{S' \sim D^m}[Q(S')]$ depends on $D$.

Fundamental tension

1. PAC-Bayes prior $P$ can depend on data distribution $D$ but cannot depend on the data $S$;
2. Our only handle on the unknown distribution $D$ is the sample $S$.

Distribution-dependent priors + KL bounds
Catoni (2004; 2007); Lever et al. (2010); Rivasplata et al. (2019)

Data-dependent priors

Use a subset of data to learn prior
Use remainder of data for bound (Ambroladze et al. 2007; Parrado-Hernández et al. 2012)

Use all the data + differential privacy (D. and Roy 2018a)
Can we exploit optimal priors?

Optimal prior $P^* = \mathbb{E}_{S' \sim D} [Q(S')]$ depends on $D$.

**Fundamental tension**

1. PAC-Bayes prior $P$ can depend on data distribution $D$ but cannot depend on the data $S$;

Distribution-dependent priors + KL bounds

Catoni (2004; 2007); Lever et al. (2010); Rivasplata et al. (2019)

Data-dependent priors

Use a subset of data to learn prior

Use remainder of data for bound (Ambroladze et al. 2007; Parrado-Hernández et al. 2012)

Use all the data + differential privacy (D. and Roy 2018a)
Can we exploit optimal priors?

Optimal prior $P^* = \mathbb{E}_{S' \sim D^m}[Q(S')]$ depends on $D$.

**Fundamental tension**

1. PAC-Bayes prior $P$ can depend on data distribution $D$ but cannot depend on the data $S$;
2. Our only handle on the unknown distribution $D$ is the sample $S$. 

---

IAS Theory of Deep Learning Workshop 2019
Can we exploit optimal priors?

Optimal prior \( P^* = \mathbb{E}_{S' \sim D^m}[Q(S')] \) depends on \( D \).

**Fundamental tension**

1. PAC-Bayes prior \( P \) **can** depend on data distribution \( D \) but **cannot** depend on the data \( S \);
2. Our only handle on the unknown distribution \( D \) is the sample \( S \).

- **Distribution-dependent priors + KL bounds**
  - Catoni (2004; 2007); Lever et al. (2010); Rivasplata et al. (2019)
Can we exploit optimal priors?

Optimal prior \( P^* = \mathbb{E}_{S' \sim D_m}[Q(S')] \) depends on \( D \).

**Fundamental tension**

1. PAC-Bayes prior \( P \) can depend on data distribution \( D \) but cannot depend on the data \( S \);
2. Our only handle on the unknown distribution \( D \) is the sample \( S \).

- **Distribution-dependent priors + KL bounds**
  - Catoni (2004; 2007); Lever et al. (2010); Rivasplata et al. (2019)

- **Data-dependent priors**
Can we exploit optimal priors?

Optimal prior $P^* = \mathbb{E}_{S' \sim D^m}[Q(S')]$ depends on $D$.

**Fundamental tension**

1. PAC-Bayes prior $P$ can depend on data distribution $D$ but cannot depend on the data $S$;
2. Our only handle on the unknown distribution $D$ is the sample $S$.

- Distribution-dependent priors + KL bounds
  Catoni (2004; 2007); Lever et al. (2010); Rivasplata et al. (2019)

- Data-dependent priors
  - Use a subset of data to learn prior
    Use remainder of data for bound (Ambroladze et al. 2007; Parrado-Hernández et al. 2012)
Can we exploit optimal priors?

Optimal prior $P^* = \mathbb{E}_{S' \sim \mathcal{D}^m}[Q(S')]$ depends on $\mathcal{D}$.

**Fundamental tension**

1. PAC-Bayes prior $P$ **can** depend on data distribution $\mathcal{D}$ but **cannot** depend on the data $S$;
2. Our only handle on the unknown distribution $\mathcal{D}$ is the sample $S$.

- **Distribution-dependent priors + KL bounds**
  Catoni (2004; 2007); Lever et al. (2010); Rivasplata et al. (2019)

- **Data-dependent priors**
  - Use a subset of data to learn prior
    Use remainder of data for bound (Ambroladze et al. 2007; Parrado-Hernández et al. 2012)
  - Use all the data + differential privacy
    (D. and Roy 2018a)
Distribution-dependent priors (Lever et al. 2010)

\[
\begin{align*}
\text{They show } & \quad \text{KL}(Q' || P') \text{ is bounded above with probability } \geq 1 - \delta, \\
& \text{satisfying } \text{KL}(Q' || P') \leq \frac{\gamma}{\sqrt{m}} \sqrt{\frac{\ln 4}{\sqrt{m}} \delta} + \frac{\gamma^2}{4m} + \frac{\ln 4}{\sqrt{m} \delta}.
\end{align*}
\]
Lever et al. 2010 study priors and posteriors of the form

\[ dP'(w) \propto \exp\{-\gamma L_D(w)\}\,dw \quad dQ'(w|S) \propto \exp\{-\gamma L_S(w)\}\,dw \]
Lever et al. 2010 study priors and posteriors of the form

\[ dP'(w) \propto \exp\{-\gamma L_D(w)\} \, dw \quad dQ'(w|S) \propto \exp\{-\gamma L_S(w)\} \, dw \]

They show \( KL(Q'||P') \) is bounded above with probability \( \geq 1 - \delta \), satisfying

\[
KL(Q'||P') \leq \gamma \sqrt{\frac{4 \sqrt{m}}{\delta}} + \frac{\gamma^2}{4m}
\]
Lever et al. 2010 study priors and posteriors of the form

\[ dP'(w) \propto \exp\{-\gamma L_D(w)\} \, dw \quad dQ'(w|S) \propto \exp\{-\gamma L_S(w)\} \, dw \]

They show \( KL(Q'||P') \) is bounded above with probability \( \geq 1 - \delta \), satisfying

\[ KL(Q'||P') \leq \frac{\gamma}{\sqrt{m}} \sqrt{\ln \frac{4\sqrt{m}}{\delta}} + \frac{\gamma^2}{4m} \]

which yields the following PAC-Bayes bound: with probability \( \geq 1 - \delta \),

\[ \Delta(L_S(Q'), L_D(Q')) \leq \frac{1}{m} \left( \frac{\gamma}{\sqrt{m}} \sqrt{\ln \frac{4\sqrt{m}}{\delta}} + \frac{\gamma^2}{4m} + \ln \frac{4\sqrt{m}}{\delta} \right) \]
Empirical evaluation of Lever et al.’s bounds

\[ dQ(w|S) \propto \exp\{-\gamma L_S(w)\}\, dw \]
Empirical evaluation of Lever et al.’s bounds

\[ dQ(w|S) \propto \exp\{-\gamma L_S(w)\} \, dw \]
Distribution-dependent approximations of optimal priors via privacy

Summary: Lever et al. bound vacuous once $\gamma$ large enough to fit random labels.

Recall: PAC-Bayes prior $P$ can depend on data distribution $D$ but not data. But data is our only handle on $D$.

Idea: If we use the data $S$ to choose a prior $P(S)$, but in a way that is stable to changes to $S$, then $P(S)$ is "almost" independent from $S$.

Theorem (D. and Roy, 2018a).

Let $P(S)$ be an $\epsilon$-differentially private prior. Then, with probability $\geq 1 - \delta$ over an i.i.d. sample $S$ from an unknown distribution,

$$(\forall Q) \Delta(L_S(Q), L_D(Q)) \leq KL(Q || P(S)) + \ln 4 \sqrt{m \delta / m + \epsilon^2 / 2 + \epsilon \sqrt{\ln 4 / \delta} / m}$$

Challenge: $\epsilon$-differential privacy for $\epsilon \ll 1$ is hard to achieve.

Solution: We show that being close in Wasserstein to a private mechanism suffices to yield a generalization bound.

See different approach based on stability by Rivasplata et al. (2018).
Distribution-dependent approximations of optimal priors via privacy

▶ Summary: Lever et al. bound vacuous once $\gamma$ large enough to fit random labels.

Recall: PAC-Bayes prior $P$ can depend on data distribution $D$ but not data. But data is our only handle on $D$.

▶ Idea: If we use the data $S$ to choose a prior $P(S)$, but in a way that is stable to changes to $S$, then $P(S)$ is “almost” independent from $S$.

Theorem (D. and Roy, 2018a). Let $P(S)$ be an $\epsilon$-differentially private prior. Then, with probability $\geq 1 - \delta$ over an i.i.d. sample $S$ from an unknown distribution, $\forall Q \quad \Delta(L_S(Q), L_D(Q)) \leq KL(Q || P(S)) + ln 4 \sqrt{m/\delta^2} + \epsilon^2/2 + \epsilon \sqrt{ln 4/\delta^2 m}$

▶ Challenge: $\epsilon$-differential privacy for $\epsilon \ll 1$ is hard to achieve.

▶ Solution: We show that being close in Wasserstein to a private mechanism suffices to yield a generalization bound.

▶ See different approach based on stability by Rivasplata et al. (2018).
Distribution-dependent approximations of optimal priors via privacy

- Summary: Lever et al. bound vacuous once $\gamma$ large enough to fit random labels.
- Recall: PAC-Bayes prior $P$ can depend on data distribution $\mathcal{D}$ but not data. But data is our only handle on $\mathcal{D}$.

Theorem (D. and Roy, 2018a). Let $P(S)$ be an $\epsilon$-differentially private prior. Then, with probability $\geq 1 - \delta$ over an i.i.d. sample $S$ from an unknown distribution, 

\[ \forall Q \] \[ \Delta(L_S(Q), L_D(Q)) \leq KL(Q \parallel P(S)) + \ln 4 \sqrt{\frac{m}{\delta}} + \epsilon^2 / 2 + \epsilon \sqrt{\ln 4 / \delta} m \]

Challenge: $\epsilon$-differential privacy for $\epsilon \ll 1$ is hard to achieve.

Solution: We show that being close in Wasserstein to a private mechanism suffices to yield a generalization bound.

See different approach based on stability by Rivasplata et al. (2018).
Distribution-dependent approximations of optimal priors via privacy

- Summary: Lever et al. bound vacuous once γ large enough to fit random labels.
- Recall: PAC-Bayes prior $P$ can depend on data distribution $\mathcal{D}$ but not data. But data is our only handle on $\mathcal{D}$.
- Idea: If we use the data $S$ to choose a prior $\mathcal{P}(S)$, but in a way that is stable to changes to $S$, then $\mathcal{P}(S)$ is “almost” independent from $S$.

**Theorem (D. and Roy, 2018a).** Let $\mathcal{P}(S)$ be an $\epsilon$-differentially private prior. Then, with probability $\geq 1 - \delta$ over an i.i.d. sample $S$ from an unknown distribution,

$$(\forall Q) \Delta(L_S(Q), L_\mathcal{D}(Q)) \leq \frac{\text{KL}(Q \| \mathcal{P}(S)) + \ln \frac{4\sqrt{m}}{\delta}}{m} + \frac{\epsilon^2}{2} + \epsilon \sqrt{\frac{\ln 4/\delta}{2m}}$$
Distribution-dependent approximations of optimal priors via privacy

- Summary: Lever et al. bound vacuous once \( \gamma \) large enough to fit random labels.
- Recall: PAC-Bayes prior \( P \) can depend on data distribution \( D \) but not data. But data is our only handle on \( D \).
- Idea: If we use the data \( S \) to choose a prior \( \mathcal{P}(S) \), but in a way that is stable to changes to \( S \), then \( \mathcal{P}(S) \) is “almost” independent from \( S \).

\textbf{Theorem (D. and Roy, 2018a).} Let \( \mathcal{P}(S) \) be an \( \epsilon \)-differentially private prior. Then, with probability \( \geq 1 - \delta \) over an i.i.d. sample \( S \) from an unknown distribution,

\[
(\forall Q) \ \Delta(L_S(Q), L_D(Q)) \leq \frac{\text{KL}(Q \| \mathcal{P}(S)) + \ln \frac{4\sqrt{m}}{\delta}}{m} + \frac{\epsilon^2}{2} + \epsilon \sqrt{\frac{\ln 4/\delta}{2m}}
\]

- Challenge: \( \epsilon \)-differential privacy for \( \epsilon \ll 1 \) is hard to achieve.
Distribution-dependent approximations of optimal priors via privacy

- Summary: Lever et al. bound vacuous once $\gamma$ large enough to fit random labels.
- Recall: PAC-Bayes prior $P$ can depend on data distribution $D$ but not data. But data is our only handle on $D$.
- Idea: If we use the data $S$ to choose a prior $P(S)$, but in a way that is stable to changes to $S$, then $P(S)$ is “almost” independent from $S$.

**Theorem (D. and Roy, 2018a).** Let $P(S)$ be an $\epsilon$-differentially private prior. Then, with probability $\geq 1 - \delta$ over an i.i.d. sample $S$ from an unknown distribution,

$$
(\forall Q) \Delta(L_S(Q), L_D(Q)) \leq \frac{\text{KL}(Q\|P(S)) + \ln \frac{4\sqrt{m}}{\delta}}{m} + \frac{\epsilon^2/2 + \epsilon \sqrt{\ln 4/\delta}}{2m}
$$

- Challenge: $\epsilon$-differential privacy for $\epsilon \ll 1$ is hard to achieve.
- Solution: We show that being close in Wasserstein to a private mechanism suffices to yield a generalization bound.
- See different approach based on stability by Rivasplata et al. (2018).
A question of interpretation

\[ \Delta \left( L_S(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S)\| P^*) + \ln I^\Delta(m)/\delta}{m}. \]

- Numerous approaches exist to approximate \( P^* = \mathbb{E}[Q(S)] \) analytically, with data, and with privacy/stability.

Theorem (D., Roy, Hsu, Gharbieh 2019+). Informally, there's a distribution, loss, and learning algorithm such that a PAC-Bayes bound with oracle prior \( P^*(S') = \mathbb{E}[Q(S)] \) is vacuous, but same bound on a subset \( S \setminus S' \) with data-dependent oracle prior \( P^*(S') = \mathbb{E}[Q(S)|S'] \) is nonvacuous.
A question of interpretation

\[ \Delta\left(L_S(Q(S)), L_D(Q(S))\right) \leq \frac{\text{KL}(Q(S)\|P^*) + \ln I^\Delta(m)/\delta}{m}. \]

- Numerous approaches exist to approximate \( P^* = \mathbb{E}[Q(S)] \) analytically, with data, and with privacy/stability.
- Is approximating \( P^* \) actually optimal?
A question of interpretation

\[ \Delta \left( L_S(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S) \| P^*) + \ln I^\Delta(m)/\delta}{m}. \]

- Numerous approaches exist to approximate \( P^* = \mathbb{E}[Q(S)] \) analytically, with data, and with privacy/stability.
- Is approximating \( P^* \) actually optimal?
A question of interpretation

\[ \Delta \left( L_S(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S) || P^*) + \ln \mathcal{I}^\Delta(|S|) / \delta}{|S|} . \]

▶ Numerous approaches exist to approximate \( P^* = \mathbb{E}[Q(S)] \) analytically, with data, and with privacy/stability.
▶ Is approximating \( P^* \) actually optimal?
A question of interpretation

\[ \Delta \left( L_{S \setminus S'}(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S)\|P^*) + \ln \mathcal{I}(S \setminus S')}{|S \setminus S'|} \cdot \frac{1}{\delta} \]

- Numerous approaches exist to approximate \( P^* = \mathbb{E}[Q(S)] \) analytically, with data, and with privacy/stability.
- Is approximating \( P^* \) actually optimal?
A question of interpretation

\[ \Delta \left( L_{S \setminus S'}(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S) \| P^*(S')) + \ln I^\Delta(|S \setminus S'|)/\delta}{|S \setminus S'|}. \]

- Numerous approaches exist to approximate \( P^* = \mathbb{E}[Q(S)] \) analytically, with data, and with privacy/stability.
- Is approximating \( P^* \) actually optimal?
A question of interpretation

\[ \Delta \left( L_{S \setminus S'}(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S) \| P^*(S')) + \ln I^\Delta(\mid S \setminus S'\mid)}{|S \setminus S'|} / \delta. \]

- Numerous approaches exist to approximate \( P^* = \mathbb{E}[Q(S)] \) analytically, with data, and with privacy/stability.
- Is approximating \( P^* \) actually optimal?

If we only use \( S \setminus S' \) to estimate generalization error then the “optimal prior” is \( P^*(S') = \mathbb{E}[Q(S) | S'] \).
A question of interpretation

\[ \Delta \left( L_{S \setminus S'}(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S) \| P^*(S')) + \ln I^\Delta(|S \setminus S'|)/\delta}{|S \setminus S'|}. \]

- Numerous approaches exist to approximate \( P^* = \mathbb{E}[Q(S)] \) analytically, with data, and with privacy/stability.
- Is approximating \( P^* \) actually optimal?

If we only use \( S \setminus S' \) to estimate generalization error then the “optimal prior” is \( P^*(S') = \mathbb{E}[Q(S)|S'] \).

Would we ever want to do this?
A question of interpretation

\[ \Delta \left( L_{S\setminus S'}(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S)\|P^*(S')) + \ln I^\Delta(|S \setminus S'|)/\delta}{|S \setminus S'|}. \]

- Numerous approaches exist to approximate \( P^* = \mathbb{E}[Q(S)] \) analytically, with data, and with privacy/stability.
- Is approximating \( P^* \) actually optimal?

If we only use \( S \setminus S' \) to estimate generalization error then the “optimal prior” is \( P^*(S') = \mathbb{E}[Q(S)|S'] \).

Would we ever want to do this? Yes.

**Theorem (D., Roy, Hsu, Gharbieh 2019+).** Informally, there’s a distribution, loss, and learning algorithm such that a PAC-Bayes bound with oracle prior \( P^*(S') = \mathbb{E}[Q(S)] \) is vacuous, but same bound on a subset \( S \setminus S' \) with data-dependent oracle prior \( P^*(S') = \mathbb{E}[Q(S)|S'] \) is nonvacuous.
Recap: Towards a nonvacuous bound on SGD

\[
\Delta \left( L_{S \setminus S'}(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S) \| P^*(S')) + \ln \mathcal{I}^\Delta(|S \setminus S'|)/\delta}{|S \setminus S'|}.
\]

**In order to relate** \( Q \) **to SGD weights** \( w_{\text{SGD}} \)...

Need posterior \( Q \) tightly concentrated around weights \( w_{\text{SGD}} \) learned by SGD.
Recap: Towards a nonvacuous bound on SGD

\[ \Delta \left( L_{S\setminus S'}(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S) \| P^*(S')) + \ln I^A(S \setminus S')/\delta}{|S \setminus S'|}. \]

- **In order to relate** \( Q \) **to SGD weights** \( w_{\text{SGD}} \)... Need posterior \( Q \) tightly concentrated around weights \( w_{\text{SGD}} \) learned by SGD.

- **In order to control KL complexity term**...
  Use some data to approximate data-dependent oracle prior \( P^*(S') = \mathbb{E}[Q(S)|S'] \).
Recap: Towards a nonvacuous bound on SGD

\[ \Delta \left( L_{S\setminus S'}(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S)\|P^*(S')) + \ln \mathcal{I}^\Delta(|S \setminus S'|)/\delta}{|S \setminus S'|}. \]

- **In order to relate** \( Q \) **to SGD weights** \( w_{\text{SGD}} \)...

  Need posterior \( Q \) tightly concentrated around weights \( w_{\text{SGD}} \) learned by SGD.

- **In order to control KL complexity term**...

  Use some data to approximate data-dependent oracle prior \( P^*(S') = \mathbb{E}[Q(S)|S'] \).
  Empirical risk term \( L_{S\setminus S'}(Q(S)) \) computed on remainder of data.
Recap: Towards a nonvacuous bound on SGD

$$\Delta \left( L_{S \setminus S'}(Q(S)), L_D(Q(S)) \right) \leq \frac{\text{KL}(Q(S)||P^*(S')) + \ln \mathcal{I}^A(|S \setminus S'|)/\delta}{|S \setminus S'|}.$$ 

- **In order to relate $Q$ to SGD weights $w_{SGD}$...**
  Need posterior $Q$ tightly concentrated around weights $w_{SGD}$ learned by SGD.

- **In order to control KL complexity term...**
  Use some data to approximate data-dependent oracle prior $P^*(S') = \mathbb{E}[Q(S)|S'].$
  Empirical risk term $L_{S \setminus S'}(Q(S))$ computed on remainder of data.

**How might we approximate $P^*(S') = \mathbb{E}[Q(S)|S']$?**
Approximating $P^*(S') = \mathbb{E}[Q(S)|S']$.

Consider a Gaussian prior $P$ and posterior $Q$:

- Fix posterior mean $w_{\text{SGD}}(S)$ to SGD weights, optimize $\text{diag}(s)$.
- We must choose prior mean $\bar{w}(S')$ before seeing full data $S$. 
Approximating $P^*(S') = \mathbb{E}[Q(S)|S']$.

Consider a Gaussian prior $P$ and posterior $Q$:

Let $Q(S) = N(w_{\text{SGD}}(S), \text{diag}(s))$ where $w_{\text{SGD}}(S)$ is the output of SGD.
Approximating $P^*(S') = \mathbb{E}[Q(S)|S']$.

Consider a Gaussian prior $P$ and posterior $Q$:

Let $Q(S) = N(w_{SGD}(S), \text{diag}(s))$ where $w_{SGD}(S)$ is the output of SGD.

Let $P(S') = N(\bar{w}(S'), \lambda_0 I)$ where $\bar{w}(S')$ is some data-dependent parameter.
Approximating $P^*(S') = \mathbb{E}[Q(S)|S']$.

Consider a Gaussian prior $P$ and posterior $Q$:

Let $Q(S) = N(w_{\text{SGD}}(S), \text{diag}(s))$ where $w_{\text{SGD}}(S)$ is the output of SGD.

Let $P(S') = N(\bar{w}(S'), \lambda_0 I)$ where $\bar{w}(S')$ is some data-dependent parameter.

$$\text{KL}(Q(S)||P(S')) = \frac{1}{2\lambda_0} \| w_{\text{SGD}}(S) - \bar{w}(S') \|^2 + \frac{1}{2} \sum_i \psi(\lambda_0, s_i).$$
Approximating $P^*(S') = \mathbb{E}[Q(S)|S']$.

Consider a Gaussian prior $P$ and posterior $Q$:

Let $Q(S) = N(w_{\text{SGD}}(S), \text{diag}(s))$ where $w_{\text{SGD}}(S)$ is the output of SGD.

Let $P(S') = N(\bar{w}(S'), \lambda_0 \mathbb{I})$ where $\bar{w}(S')$ is some data-dependent parameter.

$$\text{KL}(Q(S)||P(S')) = \frac{1}{2\lambda_0} ||w_{\text{SGD}}(S) - \bar{w}(S')||_2^2 + \frac{1}{2} \sum_i \psi(\lambda_0, s_i).$$

- Fix posterior mean $w_{\text{SGD}}(S)$ to SGD weights, optimize $\text{diag}(s)$. 
Approximating $P^*(S') = \mathbb{E}[Q(S) | S']$.

Consider a Gaussian prior $P$ and posterior $Q$:

Let $Q(S) = N(w_{SGD}(S), \text{diag}(s))$ where $w_{SGD}(S)$ is the output of SGD.

Let $P(S') = N(\bar{w}(S'), \lambda_0 I)$ where $\bar{w}(S')$ is some data-dependent parameter.

$$\text{KL}(Q(S)||P(S')) = \frac{1}{2\lambda_0} \|w_{SGD}(S) - \bar{w}(S')\|^2_2 + \frac{1}{2} \sum_i \psi(\lambda_0, s_i).$$

- Fix posterior mean $w_{SGD}(S)$ to SGD weights, optimize $\text{diag}(s)$.
- We must choose prior mean $\bar{w}(S')$ before seeing full data $S$.

$$\bar{w}(S') = \arg \min_{w'} \mathbb{E}[\|w_{SGD}(S) - w'\|^2_2 | S']$$
Approximating $P^*(S') = \mathbb{E}[Q(S)|S']$.

Consider a Gaussian prior $P$ and posterior $Q$:

Let $Q(S) = N(w_{\text{SGD}}(S), \text{diag}(s))$ where $w_{\text{SGD}}(S)$ is the output of SGD.

Let $P(S') = N(\tilde{w}(S'), \lambda_0 \mathbb{I})$ where $\tilde{w}(S')$ is some data-dependent parameter.

\[
\text{KL}(Q(S)||P(S')) = \frac{1}{2\lambda_0} \| w_{\text{SGD}}(S) - \tilde{w}(S') \|_2^2 + \frac{1}{2} \sum_i \psi(\lambda_0, s_i).
\]

- Fix posterior mean $w_{\text{SGD}}(S)$ to SGD weights, optimize $\text{diag}(s)$.
- We must choose prior mean $\tilde{w}(S')$ before seeing full data $S$.

\[
\tilde{w}(S') = \arg \min_{w'} \mathbb{E}[\| w_{\text{SGD}}(S, U) - w' \|_2^2 | S']
\]
Approximating $P^*(S') = \mathbb{E}[Q(S)|S']$. 

Consider a Gaussian prior $P$ and posterior $Q$: 

Let $Q(S) = N(w_{SGD}(S), \text{diag}(s))$ where $w_{SGD}(S)$ is the output of SGD.

Let $P(S') = N(\bar{w}(S'), \lambda_0 I)$ where $\bar{w}(S')$ is some data-dependent parameter.

$$\text{KL}(Q(S)||P(S')) = \frac{1}{2\lambda_0} \|w_{SGD}(S) - \bar{w}(S')\|_2^2 + \frac{1}{2} \sum_i \psi(\lambda_0, s_i).$$

- Fix posterior mean $w_{SGD}(S)$ to SGD weights, optimize $\text{diag}(s)$.
- We must choose prior mean $\bar{w}(S')$ before seeing full data $S$.

$$\bar{w}(S', U) = \arg \min_{w'} \mathbb{E}[\|w_{SGD}(S, U) - w'\|_2^2 | S', U]$$
Use SGD to predict SGD

\[ \tilde{w}(S', U) = \arg \min_{w'} \mathbb{E}[\|w_{\text{SGD}}(S, U) - w'\|_2^2 \mid S', U] \]
\[
\bar{w}(S', U) = \arg \min_{w'} \mathbb{E}[\|w_{\text{SGD}}(S, U) - w'\|^2_2 \mid S', U]
\]

- \(w_{\text{SGD}}(S, U)\) should be equivalent to SGD on the full data set. Since \(S' \subseteq S\) is a random subset, we’re free to choose \(S'\) to be first data processed by SGD.
Use SGD to predict SGD

\[ \bar{w}(S', U) = \arg \min_{w'} \mathbb{E}[\| w_{\text{SGD}}(S, U) - w' \|^2 \ | \ S', U] \]

- \( w_{\text{SGD}}(S, U) \) should be equivalent to SGD on the full data set. Since \( S' \subseteq S \) is a random subset, we’re free to choose \( S' \) to be first data processed by SGD.

- We will approximate \( \bar{w}(S', U) \) by running SGD on the subset \( S' \) to convergence. By design, SGD on \( S' \) will match the initial behavior of SGD on \( S \).
Example: SGD on $S'$ predicting SGD on $S$
Example: SGD on $S'$ predicting SGD on $S$
Example: SGD on $S'$ predicting SGD on $S$
Example: SGD on $S'$ predicting SGD on $S$
How well are we predicting the weights learned by SGD?

MNIST, FC (2 hidden layers).
Data-dependent oracle priors for neural networks

\[ \alpha = \frac{|S'|}{|S|} \]

\[
\text{Scaled squared L2} = \frac{\|w_{SGD} - \bar{w}\|^2}{(1 - \alpha)|S|^2}.
\]

Similar results found for networks trained on Fashion-MNIST and CIFAR10 datasets.
Coupled data-dependent approximate oracle priors and posteriors

Gaussian Lenet5 networks with **means equal to SGD trained on 30k examples from MNIST.**

![Graph](image)
Test error and PAC-Bayes generalization bounds with isotropic prior covariance. The best test error bound on MNIST, LeNet5 (approximately 11%) is significantly better than the 46% bound by Zhou et al., 2018.
Oracle access to optimal prior covariance

For a Gaussian prior $P_\Lambda$ with diagonal covariance $\Lambda = \text{diag}(\lambda_i)$, the KL term is

$$\text{KL}(Q(S)||P_\Lambda(S')) = \frac{1}{2}(w_{\text{SGD}} - \bar{w})'\Lambda(w_{\text{SGD}} - \bar{w}) + \frac{1}{2} \sum_i \Psi(\lambda_i, s_i)$$
Oracle access to optimal prior covariance

For a Gaussian prior $P_\Lambda$ with diagonal covariance $\Lambda = \text{diag}(\lambda_i)$, the KL term is

$$\text{KL}(Q(S)||P_\Lambda(S')) = \frac{1}{2}(w_{\text{SGD}} - \bar{w})'\Lambda(w_{\text{SGD}} - \bar{w}) + \frac{1}{2} \sum_i \Psi(\lambda_i, s_i)$$

- How much could an oracle estimate of $\Lambda$ help?
Oracle access to optimal prior covariance

For a Gaussian prior $P_\Lambda$ with diagonal covariance $\Lambda = \text{diag}(\lambda_i)$, the KL term is

$$\text{KL}(Q(S)||P_\Lambda(S')) = \frac{1}{2} (w_{\text{SGD}} - \bar{w})' \Lambda (w_{\text{SGD}} - \bar{w}) + \frac{1}{2} \sum_i \Psi(\lambda_i, s_i)$$

How much could an oracle estimate of $\Lambda$ help?

Optimizing the KL bound in terms of $\Lambda$, we obtain

$$\min_\Lambda \text{KL}(Q(S)||P_\Lambda(S')) = \frac{1}{2} \sum_i \ln \left( 1 + \frac{1}{s_i} (w_{\text{SGD}}^i - \bar{w}_i)^2 \right)$$
Oracle access to optimal prior covariance

For a Gaussian prior $P_{\Lambda}$ with diagonal covariance $\Lambda = \text{diag}(\lambda_i)$, the KL term is

$$\text{KL}(Q(S)\|P_{\Lambda}(S')) = \frac{1}{2} (w_{\text{SGD}} - \bar{w})' \Lambda (w_{\text{SGD}} - \bar{w}) + \frac{1}{2} \sum_i \Psi(\lambda_i, s_i)$$

- How much could an oracle estimate of $\Lambda$ help?

- Optimizing the KL bound in terms of $\Lambda$, we obtain

$$\min_{\Lambda} \text{KL}(Q(S)\|P_{\Lambda}(S')) = \frac{1}{2} \sum_i \ln \left(1 + \frac{1}{s_i} (w^i_{\text{SGD}} - \bar{w}_i)^2 \right)$$

- This bound represents the best we could hope to achieve and allows us to test limits of proposed mean prediction $\bar{w}(S', U)$. 
Gaussian network bounds with oracle data-dependent prior covariance

MNIST, Lenet-5.

The bounds are hypothetical.
Directly optimizing Variational data-dependent PAC-Bayes generalization bound.

Apply these same ideas (data-dependency and coupling) to self-bounded learning.
Recap and Conclusion

▶ Using fraction of data $S' \subseteq S$ to predict SGD on $S$ leads to significant improvement over priors centered at initialization.

▶ Data-dependence leads to predictions approximately as accurate as having fresh "ghost" samples.

▶ Theory suggests this type of data-dependent oracle prior may be necessary for tight PAC-Bayes bounds.

▶ We're still far from studying SGD itself: Stochastic neural networks in our studies were severely underfit due to looseness of the KL term during PAC-Bayes optimization. Need to understand the pareto-optimal frontier.

▶ Study of Gibbs classifiers "concentrated" near SGD weights may be a fruitful (suggestive) test bed for generalization ideas.
Using fraction of data $S' \subseteq S$ to predict SGD on $S$ leads to significant improvement over priors centered at initialization.
Recap and Conclusion

▶ Using fraction of data $S' \subseteq S$ to predict SGD on $S$ leads to significant improvement over priors centered at initialization.
  ▶ Data-dependence leads to predictions approximately as accurate as having fresh “ghost” samples.
Recap and Conclusion

- Using fraction of data $S' \subseteq S$ to predict SGD on $S$ leads to significant improvement over priors centered at initialization.
  - Data-dependence leads to predictions approximately as accurate as having fresh “ghost” samples.
- Theory suggests this type of data-dependent oracle prior may be necessary for tight PAC-Bayes bounds.
Recap and Conclusion

- Using fraction of data $S' \subseteq S$ to predict SGD on $S$ leads to significant improvement over priors centered at initialization.
  - Data-dependence leads to predictions approximately as accurate as having fresh “ghost” samples.
- Theory suggests this type of data-dependent oracle prior may be necessary for tight PAC-Bayes bounds.
- We’re still far from studying SGD itself: Stochastic neural networks in our studies were severely underfit due to looseness of the KL term during PAC-Bayes optimization. Need to understand the pareto-optimal frontier.
Recap and Conclusion

- Using fraction of data $S' \subseteq S$ to predict SGD on $S$ leads to significant improvement over priors centered at initialization.
  - Data-dependence leads to predictions approximately as accurate as having fresh “ghost” samples.
- Theory suggests this type of data-dependent oracle prior may be necessary for tight PAC-Bayes bounds.
- We’re still far from studying SGD itself: Stochastic neural networks in our studies were severely underfit due to looseness of the KL term during PAC-Bayes optimization. Need to understand the pareto-optimal frontier.
- Study of Gibbs classifiers “concentrated” near SGD weights may be a fruitful (suggestive) test bed for generalization ideas.