Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates

Jeffrey NEGREA — University of Toronto; Vector Institute
Mahdi HAGHIFAM — University of Toronto
Gintarė Karolina DŽIUGAITĖ — Element AI
Ashish KHISTI — University of Toronto
Daniel M. ROY — University of Toronto; Vector Institute

Stochastic Gradient Langevin Dynamics (SGLD)

Let $S_0, S_1, S_2, \ldots$ be independent random subsets of S of size $b_0, b_1, \ldots$.

Raginsky, Rakhlin, Telgarsky 17 gave nonasymptotic risk bounds for SGLD:

$$W_{t+1} = W_t - \eta_t \nabla \tilde{L}_{S_t}(W_t) + \sqrt{2\eta_t/\beta_t}\, \varepsilon_t.$$

where

- $\varepsilon_t \sim \mathcal{N}(0, \mathbb{I}_d)$ i.i.d.,
- η_t is learning rate,
- β_t is inverse temperature,
- $\tilde{L}_S(w) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(w, z_i)$.

Stochastic Gradient Langevin Dynamics (SGLD)

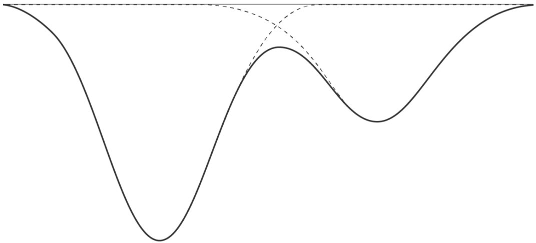
Let $S_0, S_1, S_2, \ldots$ be independent random subsets of S of size $b_0, b_1, \ldots$.

Raginsky, Rakhlin, Telgarsky 17 gave nonasymptotic risk bounds for SGLD:

$$W_{t+1} = W_t - \eta_t \nabla \tilde{L}_{S_t}(W_t) + \sqrt{2\eta_t/\beta_t}\, \varepsilon_t.$$

where

- $\varepsilon_t \sim \mathcal{N}(0, \mathbb{I}_d)$ i.i.d.,
- η_t is learning rate,
- β_t is inverse temperature,
- $\tilde{L}_S(w) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(w, z_i)$.



Stochastic Gradient Langevin Dynamics (SGLD)

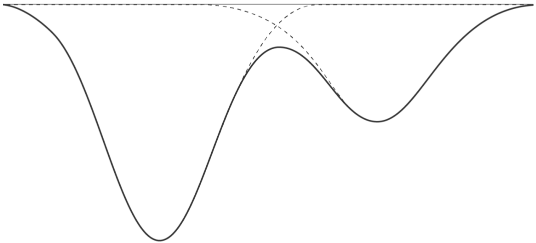
Let $S_0, S_1, S_2, \dots$ be independent random subsets of S of size $b_0, b_1, \dots$.

Raginsky, Rakhlin, Telgarsky 17 gave nonasymptotic risk bounds for SGLD:

$$W_{t+1} = W_t - \eta_t \nabla \tilde{L}_{S_t}(W_t) + \sqrt{2\eta_t/\beta_t}\,\varepsilon_t.$$

where

- $\varepsilon_t \sim \mathcal{N}(0, \mathbb{I}_d)$ i.i.d.,
- η_t is learning rate,
- β_t is inverse temperature,
- $\tilde{L}_S(w) = \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(w, z_i)$.



This talk: building on sequential analysis of Pensia, Jog, and Loh (2017).

Mutual information bound

Let $S = (Z_1, \ldots, Z_n) \sim \mathcal{D}^n$ and W be learned weights (i.e., a random element in $\mathbb{R}^d$).

$$\mathrm{EGE}(W, S) = \mathbb{E}[L_{\mathcal{D}}(W) - L_S(W)]$$

Suppose $\ell(Z_1, w)$ is σ-sub-Gaussian for every $w \in \mathbb{R}^d$.

Theorem (XR17, RZ15). $|\mathrm{EGE}(W, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W; S)}{|S|}}$.

Mutual information bound

Let $S = (Z_1, \ldots, Z_n) \sim \mathcal{D}^n$ and W be learned weights (i.e., a random element in $\mathbb{R}^d$).

$$\mathrm{EGE}(W, S) = \mathbb{E}[L_{\mathcal{D}}(W) - L_S(W)]$$

Suppose $\ell(Z_1, w)$ is σ-sub-Gaussian for every $w \in \mathbb{R}^d$.

Theorem (XR17, RZ15). $|\mathrm{EGE}(W, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W; S)}{|S|}}.$

- Does this theorem "explain" SGLD generalization?

Mutual information bound

Let $S = (Z_1, \dots, Z_n) \sim \mathcal{D}^n$ and W be learned weights (i.e., a random element in $\mathbb{R}^d$).

$$\mathrm{EGE}(W, S) = \mathbb{E}[L_{\mathcal{D}}(W) - L_S(W)]$$

Suppose $\ell(Z_1, w)$ is σ-sub-Gaussian for every $w \in \mathbb{R}^d$. Let $Q_T(S) = \mathbb{P}^S[W_T]$ and $P_T = \mathbb{P}[W_T]$.

Theorem (XR17, RZ15). $|\mathrm{EGE}(W, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W; S)}{|S|}}$.

- Does this theorem "explain" SGLD generalization?

Mutual information bound

Let $S = (Z_1, \ldots, Z_n) \sim \mathcal{D}^n$ and W be learned weights (i.e., a random element in $\mathbb{R}^d$).

$$\mathrm{EGE}(W, S) = \mathbb{E}[L_{\mathcal{D}}(W) - L_S(W)]$$

Suppose $\ell(Z_1, w)$ is σ-sub-Gaussian for every $w \in \mathbb{R}^d$. Let $Q_T(S) = \mathbb{P}^S[W_T]$ and $P_T = \mathbb{P}[W_T]$.

Theorem (XR17, RZ15). $|\mathrm{EGE}(W, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W; S)}{|S|}} = \sqrt{2\sigma^2 \frac{\mathbb{E}[\mathrm{KL}(Q_T(S) || P_T)]}{|S|}}.$

- Does this theorem "explain" SGLD generalization?

Mutual information bound

Let $S = (Z_1, \ldots, Z_n) \sim \mathcal{D}^n$ and W be learned weights (i.e., a random element in $\mathbb{R}^d$).

$$\mathrm{EGE}(W, S) = \mathbb{E}[L_{\mathcal{D}}(W) - L_S(W)]$$

Suppose $\ell(Z_1, w)$ is σ-sub-Gaussian for every $w \in \mathbb{R}^d$. Let $Q_T(S) = \mathbb{P}^S[W_T]$ and $P_T = \mathbb{P}[W_T]$.

Theorem (XR17, RZ15). $|\mathrm{EGE}(W, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W; S)}{|S|}} = \sqrt{2\sigma^2 \frac{\mathbb{E}[\mathrm{KL}(Q_T(S) || P_T)]}{|S|}}$.

- Does this theorem "explain" SGLD generalization?
- Proof (via Donsker–Varadhan) suggests which procedures enjoy tightest bounds.

Mutual information bound

Let $S = (Z_1, \ldots, Z_n) \sim \mathcal{D}^n$ and W be learned weights (i.e., a random element in $\mathbb{R}^d$).

$$\mathrm{EGE}(W, S) = \mathbb{E}[L_{\mathcal{D}}(W) - L_S(W)]$$

Suppose $\ell(Z_1, w)$ is σ-sub-Gaussian for every $w \in \mathbb{R}^d$. Let $Q_T(S) = \mathbb{P}^S[W_T]$ and $P_T = \mathbb{P}[W_T]$.

Theorem (XR17, RZ15). $|\mathrm{EGE}(W, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W; S)}{|S|}} = \sqrt{2\sigma^2 \frac{\mathbb{E}[\mathrm{KL}(Q_T(S) || P_T)]}{|S|}}$.

- Does this theorem "explain" SGLD generalization?
- Proof (via Donsker–Varadhan) suggests which procedures enjoy tightest bounds.
- Statistical barrier: $\mathrm{I}(W; S)$ depends on unknown $\mathcal{D}$.

Mutual information bound

Let $S = (Z_1, \ldots, Z_n) \sim \mathcal{D}^n$ and W be learned weights (i.e., a random element in $\mathbb{R}^d$).

$$\mathrm{EGE}(W, S) = \mathbb{E}[L_{\mathcal{D}}(W) - L_S(W)]$$

Suppose $\ell(Z_1, w)$ is σ-sub-Gaussian for every $w \in \mathbb{R}^d$. Let $Q_T(S) = \mathbb{P}^S[W_T]$ and $P_T = \mathbb{P}[W_T]$.

Theorem (XR17, RZ15). $|\mathrm{EGE}(W, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W; S)}{|S|}} = \sqrt{2\sigma^2 \frac{\mathbb{E}[\mathrm{KL}(Q_T(S)||P_T)]}{|S|}}$.

- Does this theorem "explain" SGLD generalization?
- Proof (via Donsker–Varadhan) suggests which procedures enjoy tightest bounds.
- Statistical barrier: $\mathrm{I}(W; S)$ depends on unknown $\mathcal{D}$.
- Computational barrier: even if $\mathcal{D}$ were known, $\mathbb{P}[W]$ often intractable.

Bounding $\mathrm{I}(S; W_T)$ for SGLD

Let $S_0, S_1, S_2, \ldots$ be random minibatches of S. The iterates $W_0, W_1, \ldots, W_T$ of SGLD satisfy

$$W_{i+1} = W_i - \eta_i \nabla \tilde{L}_{S_i}(W_i) + \sqrt{2\eta_i/\beta_i}\,\varepsilon_i.$$

Theorem (XR17, RZ15). $|\mathrm{EGE}(W_T, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(S; W_T)}{|S|}}$.

Even if $\mathcal{D}$ were known, $\mathrm{I}(W_T; S)$ involves intractable marginal $\mathbb{P}W_T$.

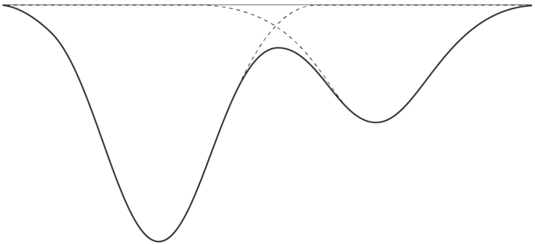
Bounding $\mathrm{I}(S; W_T)$ for SGLD

Let $S_0, S_1, S_2, \ldots$ be random minibatches of S. The iterates $W_0, W_1, \ldots, W_T$ of SGLD satisfy

$$W_{i+1} = W_i - \eta_i \nabla \tilde{L}_{S_i}(W_i) + \sqrt{2\eta_i/\beta_i}\,\varepsilon_i.$$

Theorem (XR17, RZ15). $|\mathrm{EGE}(W_T, S)| \le \sqrt{2\sigma^2 \frac{\mathrm{I}(S; W_T)}{|S|}}$.

Even if $\mathcal{D}$ were known, $\mathrm{I}(W_T; S)$ involves intractable marginal $\mathbb{P}W_T$.



Bounding $\mathrm{I}(S; W_T)$ for SGLD

Let $S_0, S_1, S_2, \ldots$ be random minibatches of S. The iterates $W_0, W_1, \ldots, W_T$ of SGLD satisfy

$$W_{i+1} = W_i - \eta_i \nabla \tilde{L}_{S_i}(W_i) + \sqrt{2\eta_i/\beta_i}\, \varepsilon_i.$$

Theorem (XR17, RZ15). $|\mathrm{EGE}(W_T, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(S; W_T)}{|S|}}$.

Even if $\mathcal{D}$ were known, $\mathrm{I}(W_T; S)$ involves intractable marginal $\mathbb{P}W_T$.

[PJL18] $\quad \mathrm{I}(S; W_T) \leq \mathrm{I}(S; W_{1:T})$

Bounding $\mathrm{I}(S; W_T)$ for SGLD

Let $S_0, S_1, S_2, \ldots$ be random minibatches of S. The iterates $W_0, W_1, \ldots, W_T$ of SGLD satisfy

$$W_{i+1} = W_i - \eta_i \nabla \tilde{L}_{S_i}(W_i) + \sqrt{2\eta_i/\beta_i}\,\varepsilon_i.$$

Theorem (XR17, RZ15). $|\mathrm{EGE}(W_T, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(S; W_T)}{|S|}}$.

Even if $\mathcal{D}$ were known, $\mathrm{I}(W_T; S)$ involves intractable marginal $\mathbb{P}W_T$.

[PJL18]
$$\begin{aligned}
\mathrm{I}(S; W_T) &\leq \mathrm{I}(S; W_{1:T}) \\
&\leq \mathrm{I}(S_0; W_1) + \mathrm{I}(S_1; W_2 | W_1) \\
&\qquad + \mathrm{I}(S_2; W_3 | W_1, W_2) + \cdots + \mathrm{I}(S_{T-1}; W_T | W_{1:T-1})
\end{aligned}$$

Bounding $\mathrm{I}(S; W_T)$ for SGLD

Let $S_0, S_1, S_2, \ldots$ be random minibatches of S. The iterates $W_0, W_1, \ldots, W_T$ of SGLD satisfy

$$W_{i+1} = W_i - \eta_i \nabla \tilde{L}_{S_i}(W_i) + \sqrt{2\eta_i/\beta_i}\, \varepsilon_i.$$

Theorem (XR17, RZ15). $|\mathrm{EGE}(W_T, S)| \le \sqrt{2\sigma^2 \frac{\mathrm{I}(S; W_T)}{|S|}}$.

Even if $\mathcal{D}$ were known, $\mathrm{I}(W_T; S)$ involves intractable marginal $\mathbb{P}W_T$.

[PJL18]

$$\begin{aligned}\mathrm{I}(S; W_T) &\le \mathrm{I}(S; W_{1:T}) \\ &\le \mathrm{I}(S_0; W_1) + \mathrm{I}(S_1; W_2|W_1) \\ &\quad + \mathrm{I}(S_2; W_3|W_1, W_2) + \cdots + \mathrm{I}(S_{T-1}; W_T|W_{1:T-1})\end{aligned}$$

$$\mathrm{I}(S_{T-1}; W_T|W_{1:T-1}) \le \frac{d}{2} \ln\left(1 + \frac{\eta_i \beta_i L^2}{2d}\right) \le \eta_i \beta_i L^2/4 \quad \text{where} \quad \sup_i \|\nabla \tilde{L}_{S_i}(W_i)\|_2 \le L \text{ a.s.}$$

Bounding $\mathrm{I}(S; W_T)$ for SGLD

Let $S_0, S_1, S_2, \ldots$ be random minibatches of S. The iterates $W_0, W_1, \ldots, W_T$ of SGLD satisfy

$$W_{i+1} = W_i - \eta_i \nabla \tilde{L}_{S_i}(W_i) + \sqrt{2\eta_i/\beta_i}\, \varepsilon_i.$$

Theorem (XR17, RZ15). $|\mathrm{EGE}(W_T, S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(S; W_T)}{|S|}} \overset{\text{[PJL18]}}{\leq} \sqrt{\frac{\sigma^2}{2n} \sum_{i=1}^{T} \eta_i \beta_i L^2}\,.$

Even if $\mathcal{D}$ *were known*, $\mathrm{I}(W_T; S)$ involves intractable marginal $\mathbb{P}W_T$.

$$\text{[PJL18]} \qquad \begin{aligned} \mathrm{I}(S; W_T) &\leq \mathrm{I}(S; W_{1:T}) \\ &\leq \mathrm{I}(S_0; W_1) + \mathrm{I}(S_1; W_2 | W_1) \\ &\qquad + \mathrm{I}(S_2; W_3 | W_1, W_2) + \cdots + \mathrm{I}(S_{T-1}; W_T | W_{1:T-1}) \end{aligned}$$

$$\mathrm{I}(S_{T-1}; W_T | W_{1:T-1}) \leq \frac{d}{2} \ln\left(1 + \frac{\eta_i \beta_i L^2}{2d}\right) \leq \eta_i \beta_i L^2/4 \quad \text{where} \quad \sup_i \|\nabla \tilde{L}_{S_i}(W_i)\|_2 \leq L \text{ a.s.}$$

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W, S)| \le \sqrt{2\sigma^2 \frac{\mathrm{I}(W; S)}{|S|}}.$

Let $J \subseteq [n] = \{1, \ldots, n\}$ be uniformly distributed subset of size $|J| = m \le n$ and $J \perp\!\!\!\perp (S, W)$.
Let $S = (S_J, S_{\bar{J}})$.

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W,S)| \le \sqrt{2\sigma^2 \frac{\mathrm{I}(W;S)}{|S|}}$.

Let $J \subseteq [n] = \{1,\dots,n\}$ be uniformly distributed subset of size $|J| = m \le n$ and $J \perp\!\!\!\perp (S,W)$.
Let $S = (S_J, S_{\bar{J}})$.

Theorem (RRTWX16, BZV19, NHDKR19).

$$|\mathrm{EGE}(W,S)| \le \mathbb{E}\left[\sqrt{2\sigma^2 \frac{\mathrm{I}(W;S_J|S_{\bar{J}})}{|S_J|}}\right].$$

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W,S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W;S)}{|S|}}.$

Let $J \subseteq [n] = \{1,\ldots,n\}$ be uniformly distributed subset of size $|J| = m \leq n$ and $J \perp\!\!\!\perp (S,W)$.
Let $S = (S_J, S_{\bar{J}})$.

Note $\mathrm{I}(W;S_{\bar{J}}|S_J) = \mathbb{E}[\mathrm{I}^{S_J}(W;S_{\bar{J}})]$.

Theorem (RRTWX16, BZV19, NHDKR19).

$$|\mathrm{EGE}(W,S)| \leq \mathbb{E}\left[\sqrt{2\sigma^2 \frac{\mathrm{I}(W;S_J|S_{\bar{J}})}{|S_J|}}\right].$$

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W,S)| \le \sqrt{2\sigma^2 \frac{\mathrm{I}(W;S)}{|S|}}.$

Let $J \subseteq [n] = \{1,\ldots,n\}$ be uniformly distributed subset of size $|J| = m \le n$ and $J \perp\!\!\!\perp (S,W)$. Let $S = (S_J, S_{\bar{J}})$.

Note $\mathrm{I}(W; S_{\bar{J}} | S_J) = \mathbb{E}[\mathrm{I}^{S_J}(W; S_{\bar{J}})]$.

Theorem (NHDKR19).

$$|\mathrm{EGE}(W,S)| \le \mathbb{E}\left[\sqrt{2\sigma^2 \frac{\mathbb{E}[\mathrm{I}^{S_J}(S_{\bar{J}}; W)]}{|S_J|}}\right].$$

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W,S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W;S)}{|S|}}.$

Let $J \subseteq [n] = \{1,\ldots,n\}$ be uniformly distributed subset of size $|J| = m \leq n$ and $J \perp\!\!\!\perp (S,W)$.
Let $S = (S_J, S_{\bar{J}})$.

Note $\mathrm{I}(W; S_{\bar{J}} | S_J) = \mathbb{E}[\mathrm{I}^{S_J}(W; S_{\bar{J}})]$.

Theorem (NHDKR19).

$$|\mathrm{EGE}(W,S)| \leq \qquad \leq \mathbb{E}\left[\sqrt{2\sigma^2 \frac{\mathbb{E}[\mathrm{I}^{S_J}(S_{\bar{J}}; W)]}{|S_J|}}\right].$$

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W,S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W;S)}{|S|}}.$

Let $J \subseteq [n] = \{1,\ldots,n\}$ be uniformly distributed subset of size $|J| = m \leq n$ and $J \perp\!\!\!\perp (S,W)$.
Let $S = (S_J, S_{\bar{J}})$.

Note $\mathrm{I}(W; S_{\bar{J}}|S_J) = \mathbb{E}[\mathrm{I}^{S_J}(W; S_{\bar{J}})]$.

Theorem (NHDKR19).

$$|\mathrm{EGE}(W,S)| \leq \mathbb{E}\left[\sqrt{2\sigma^2 \frac{\mathrm{I}^{S_J}(S_{\bar{J}}; W)}{|S_J|}}\,\right] \leq \mathbb{E}\left[\sqrt{2\sigma^2 \frac{\mathbb{E}[\mathrm{I}^{S_J}(S_{\bar{J}}; W)]}{|S_J|}}\right].$$

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W,S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W;S)}{|S|}}$.

Let $J \subseteq [n] = \{1,\ldots,n\}$ be uniformly distributed subset of size $|J| = m \leq n$ and $J \perp\!\!\!\perp (S,W)$.
Let $S = (S_J, S_{\bar{J}})$.

Note $\mathrm{I}(W; S_{\bar{J}}|S_J) = \mathbb{E}[\mathrm{I}^{S_J}(W; S_{\bar{J}})]$.

Theorem (NHDKR19). *Assuming* $|J| = 1$ *and* $\ell \in [0,1]$.

$$|\mathrm{EGE}(W,S)| \leq \mathbb{E}\left[\sqrt{\mathrm{I}^{S_J}(S_{\bar{J}}; W)/2}\right]$$

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W,S)| \le \sqrt{2\sigma^2 \frac{\mathrm{I}(W;S)}{|S|}}.$

Let $J \subseteq [n] = \{1,\ldots,n\}$ be uniformly distributed subset of size $|J| = m \le n$ and $J \perp\!\!\!\perp (S,W)$.
Let $S = (S_J, S_{\bar{J}})$.

Note $\mathrm{I}(W; S_{\bar{J}} | S_J) = \mathbb{E}[\mathrm{I}^{S_J}(W; S_{\bar{J}})]$.
Let $Q(S) = \mathbb{P}^S[W]$ and $P(S_J) = \mathbb{P}^{S_J}[W]$. Then $\mathrm{I}^{S_J}(W; S_{\bar{J}}) = \mathbb{E}^{S_J}[\mathrm{KL}(Q(S)||P(S_J))]$.

Theorem (NHDKR19). *Assuming* $|J| = 1$ *and* $\ell \in [0,1]$.

$$|\mathrm{EGE}(W,S)| \le \mathbb{E}\left[\sqrt{\mathrm{I}^{S_J}(S_{\bar{J}}; W)/2}\right]$$

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W,S)| \le \sqrt{2\sigma^2 \frac{\mathrm{I}(W;S)}{|S|}}.$

Let $J \subseteq [n] = \{1,\dots,n\}$ be uniformly distributed subset of size $|J| = m \le n$ and $J \perp\!\!\!\perp (S,W)$.
Let $S = (S_J, S_{\bar{J}})$.

Note $\mathrm{I}(W; S_{\bar{J}} | S_J) = \mathbb{E}[\mathrm{I}^{S_J}(W; S_{\bar{J}})]$.
Let $Q(S) = \mathbb{P}^S[W]$ and $P(S_J) = \mathbb{P}^{S_J}[W]$. Then $\mathrm{I}^{S_J}(W; S_{\bar{J}}) = \mathbb{E}^{S_J}[\mathrm{KL}(Q(S)||P(S_J))]$.

Theorem (NHDKR19). *Assuming* $|J| = 1$ *and* $\ell \in [0,1]$.

$$|\mathrm{EGE}(W,S)| \le \mathbb{E}\left[\sqrt{\mathrm{KL}(Q(S)||P(S_J))/2}\,\right]$$

Mutual information bounds

Theorem (XR17, RZ15). $|\mathrm{EGE}(W,S)| \leq \sqrt{2\sigma^2 \frac{\mathrm{I}(W;S)}{|S|}}$.

Let $J \subseteq [n] = \{1,\dots,n\}$ be uniformly distributed subset of size $|J| = m \leq n$ and $J \perp\!\!\!\perp (S,W)$.
Let $S = (S_J, S_{\bar{J}})$.

Note $\mathrm{I}(W; S_{\bar{J}} | S_J) = \mathbb{E}[\mathrm{I}^{S_J}(W; S_{\bar{J}})]$.
Let $Q(S) = \mathbb{P}^S[W]$ and $P(S_J) = \mathbb{P}^{S_J}[W]$. Then $\mathrm{I}^{S_J}(W; S_{\bar{J}}) = \mathbb{E}^{S_J}[\mathrm{KL}(Q(S) || P(S_J))]$.

Theorem (NHDKR19). *Assuming* $|J| = 1$ *and* $\ell \in [0,1]$.

$$|\mathrm{EGE}(W,S)| \leq \mathbb{E}\left[\sqrt{\mathrm{KL}(Q(S)||P(S_J))/2}\,\right]$$ *holds for all kernels* $P(\cdot)$*!*

Chain rule for KL

Let $(W_0, W_1, \dots W_T)$ be iterates.

Define $Q = Q(S) = \mathbb{P}^S[W_{0:T}]$, $Q_t = \mathbb{P}^S[W_t]$ and $Q_{t|} = \mathbb{P}^{S,W_{0:t-1}}[W_t]$.

Let $P, P_t, P_{t|}$ be arbitrary but depending on S_J not S.

Assume $P_0 = Q_0$. Then $$\mathrm{KL}(Q_T||P_T) \leq \mathrm{KL}(Q||P) = \sum_{t=1}^{T} \mathbb{E}^{W_{0:t-1}}\mathrm{KL}(Q_{t|}||P_{t|}).$$

Data-dependent priors for full-batch SGLD (i.e., Langevin algorithm)

Let S_J be a random subset of S, of size m, chosen independently from $W_0, W_1, \ldots$. The one-step distribution $Q_{t|}$ satisfies

$$Q_{t|} = Q_{t|}(S) = \mathcal{N}\left(W_t - \eta_t \nabla \tilde{L}_S(W_t),\ 2\frac{\eta_t}{\beta_t}\mathbb{I}_d\right).$$

Consider the data-dependent prior, P,

$$P_{t|} = P_{t|}(S_J) \equiv \mathcal{N}\left(W_t - \eta_t \nabla \tilde{L}_{S_J}(W_t),\ 2\frac{\eta_t}{\beta_t}\mathbb{I}_d\right).$$

The one-step KL divergence is then

$$\mathrm{KL}(Q_{t+1|}||P_{t+1|}) = \frac{\beta_t \eta_t}{8}\|\xi_{t,J}\|_2^2 \qquad \text{where } \xi_{t,J} = \underbrace{\nabla \tilde{L}_S(W_t) - \nabla \tilde{L}_{S_J}(W_t)}_{\text{“incoherence”}}.$$

EGE bounds for SGLD

$$\xi_{t,i} = \nabla \tilde{L}_S(W_t) - \nabla \tilde{L}_{S\setminus\{i\}}(W_t)$$

Theorem (NHDKR19).

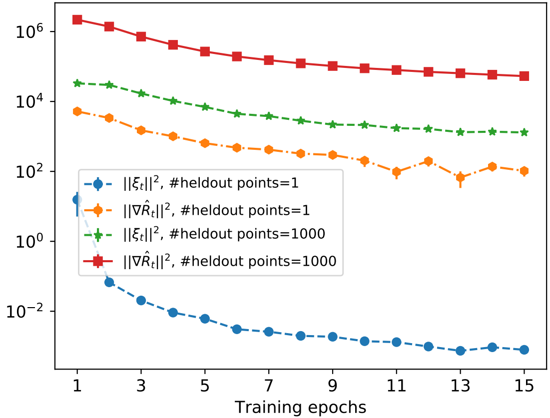
$$\mathrm{EGE}(W_T, S) \leq \mathbb{E}\sqrt{\frac{\beta\, n}{16(n-1)^2}\sum_{t=1}^{T}\eta_t \mathbb{E}^S\left[\frac{1}{n}\sum_{j=1}^{n}\|\xi_{t,j}\|^2\right]}$$

Theorem (MWZZ17).

$$\mathrm{EGE}(W_T, S) \leq \sqrt{\frac{\beta}{n}\sum_{t=1}^{T}\eta_t \mathbb{E}[\|\nabla \tilde{L}_S(W_t)\|^2]}$$

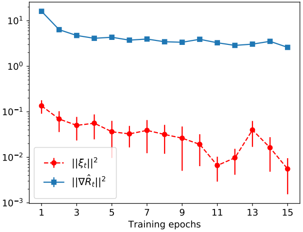
Effect of number of held-out points

MNIST, FC

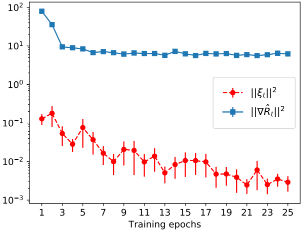


Empirical Evaluation

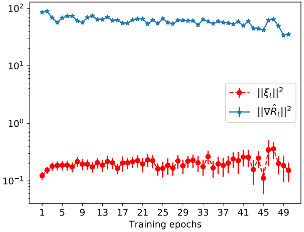
MNIST, CNN



Fashion-MNIST, CNN

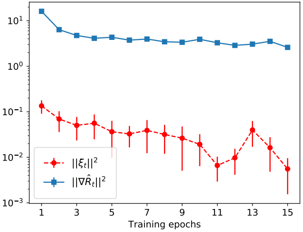


CIFAR10, CNN

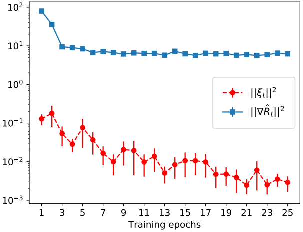


Empirical Evaluation

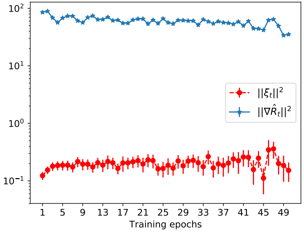
MNIST, CNN



Fashion-MNIST, CNN



CIFAR10, CNN



	MNIST with MLP			**MNIST with CNN**		
	Epoch 1	Epoch 2	Epoch 3	Epoch 1	Epoch 2	Epoch 3
Training Classification Error	$25.52 \pm 0.08\%$	$16.17 \pm 0.04\%$	$12.38 \pm 0.02\%$	$21.89 \pm 0.21\%$	$14.07 \pm 0.14\%$	$10.78 \pm 0.10\%$
Test Classification Error	$25.57 \pm 0.06\%$	$16.29 \pm 0.04\%$	$12.45 \pm 0.02\%$	$22.93 \pm 0.20\%$	$14.72 \pm 0.14\%$	$11.24 \pm 0.09\%$
Generalization Gap (Mou et al.)	$33.8 \pm 1.4\%$	$76.0 \pm 3.0\%$	$139.4 \pm 5.9\%$	$46.5 \pm 2.2\%$	$78.6 \pm 3.0\%$	$130.6 \pm 4.6\%$
Generalization Gap (Our Bound)	$10.0 \pm 1.6\%$	$20.5 \pm 4.0\%$	$29.0 \pm 6.7\%$	$15.3 \pm 2.8\%$	$25.8 \pm 4.4\%$	$49.2 \pm 10.4\%$

Conclusion

- Mutual information bounds on expected generalization [XR17,RS15]

Conclusion

- Mutual information bounds on expected generalization [XR17,RS15]
- Sequential decomposition of mutual information for SGLD [PJL18,BZV19]

Conclusion

- Mutual information bounds on expected generalization [XR17,RS15]
- Sequential decomposition of mutual information for SGLD [PJL18,BZV19]
- Vacuousness in standard regimes

Conclusion

- Mutual information bounds on expected generalization [XR17,RS15]
- Sequential decomposition of mutual information for SGLD [PJL18,BZV19]
- Vacuousness in standard regimes
- Distribution-dependence via data-dependent priors

Conclusion

- Mutual information bounds on expected generalization [XR17,RS15]
- Sequential decomposition of mutual information for SGLD [PJL18,BZV19]
- Vacuousness in standard regimes
- Distribution-dependence via data-dependent priors
- Loose but nonvacuous bounds possible empirically

Conclusion

- Mutual information bounds on expected generalization [XR17,RS15]
- Sequential decomposition of mutual information for SGLD [PJL18,BZV19]
- Vacuousness in standard regimes
- Distribution-dependence via data-dependent priors
- Loose but nonvacuous bounds possible empirically
- More work needed to understand limits of mutual information based approaches